HAL
open science

Potentiel d'un couplage entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale/forte résolution spatiale pour la sélection variétale

Maxime Ryckewaert

To cite this version:

Maxime Ryckewaert. Potentiel d'un couplage entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale/forte résolution spatiale pour la sélection variétale. Agronomie. Montpellier SupAgro, 2019. Français. NNT : . tel-03248085

HAL Id: tel-03248085

https://theses.hal.science/tel-03248085

Submitted on 3 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Génie des procédés

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau
Portée par l'Université de Montpellier

Unité de recherche UMR ITAP – Information, Technologie, Analyse environnementale, Procédés agricoles

Potentiel d'un couplage entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale/forte résolution spatiale pour la sélection variétale

Présentée par Maxime RYCKEWAERT
Le 07 Novembre 2019

Sous la direction de Jean-Michel ROGER

Devant le jury composé de

Carole SINFORT, Directrice déléguée, Montpellier SupAgro, France	Présidente du jury
Cyril RUCKEBUSCH, Professeur, Université Lille 1, France	Rapporteur
David ROUSSEAU, Professeur, Université d'Angers, France	Rapporteur
Anna DE JUAN CAPDEVILA, Professeur, Université de Barcelone, Espagne	Examinateur
Piet REYNS, Manager NIR/Research automation group, Limagrain Europe, Belgique	Examinateur
Cindy TORRES, Chef de projet traitement d'images, Vilmorin, France	Examinateur
Jean-Michel ROGER, Directeur de recherche, Irstea, France	Directeur de thèse
Nathalie GORRETTA, Ingénieur de recherche, Irstea, France	Encadrant de thèse
Fabienne HENRIOT, Responsable Abiotic Traits & Digital Phenotyping, Limagrain Europe, France	Encadrant de thèse

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

Remerciements

Cette thèse s'est déroulée dans le centre de recherche d'Irstea dans l'UMR ITAP à Montpellier, et côté industriel avec l'entreprise Limagrain Europe à Chappes. Cette thèse bénéficie de la Convention Industrielle de Formation par la Recherche (CIFRE) attribuée par l'Association Nationale de la Recherche et de la Technologie (ANRT) et est labellisé DigitAg. Je tiens à remercier l'ensemble des personnes qui ont aidé à construire ce projet de thèse. Parmi ces personnes, je remercie particulièrement mes encadrants de proximité : Alexia Gobrecht, Nathalie Gorretta, Fabienne Henriot et Jean-Michel Roger. J'ai pu bénéficier d'un encadrement pluridisciplinaire pour débuter cette thèse dans de bonnes conditions.

Je remercie Cyril Ruckebusch et David Rousseau d'avoir accepté d'être rapporteurs de cette thèse. Je remercie également les membres du jury de la soutenance : Carole Sinfort, Anna De-Juan, Piet Reyns et Cindy Torres. Je remercie les membres du comité de suivi de thèse Bruno Tisseyre, Jocelyn Chanussot, Jean-Luc Regnard, Eva Ampe et Jérémy Derory qui ont contribué énormément au développement du sujet suite aux nombreuses discussions.

Je remercie l'entreprise Limagrain ainsi que l'ensemble des personnes impliquées dans la mise en place, l'organisation et le suivi des campagnes expérimentales allant du vol de drone à la récupération de données agronomiques : Jérémy, Jérôme, Fabienne (une seconde fois), Manuel, Richard et Adrien.

Je tiens à remercier le chef d'équipe COMIC, Ryad Bendoula pour m'avoir soutenu dans les moments difficiles notamment lors de la dernière année de la thèse dans des épreuves tant redoutées telles que la rédaction et la préparation à la soutenance. Je remercie l'assistant-ingénieur Daniel Moura de m'avoir soutenu tout au long de la thèse mais aussi pour m'avoir encadré et formé sur de nombreux aspects liés à la mesure. Il a apporté énormément à la conception, au développement et à la validation du module AIRSTRIP dont les résultats, disponibles en annexe, consolident les différents blocs que forment les chapitres de cette thèse. Je remercie énormément Daphné et Anis pour le soutien moral et tous les moments de partage. Merci à Julien et Adel d'avoir été des amis et des co-bureaux exemplaires. Anice, je te remercie pour ces moments sincères. Je remercie les tenanciers de ChemHouse, Belal et Maxime pour l'ensemble des discussions chimiométriques et notamment sur les discussions portant sur le fameux pic de rien. Je remercie Arnaud, Florent et Sylvia d'apporter leurs soutiens sur cette fin de thèse.

Je remercie également l'ensemble des personnes de l'UMR ITAP : Tewfik, Jean-Paul, Jean-Philippe, Jean-François, Vincent, Sébastien, Adrien, Maud, Mathilde, Yoann, Nina, Frédéric, David, Christophe, Véronique, Gérard et Cyril. Je remercie spécialement Xavier pour tous les cafés offerts et les discussions matinales de 7h30.

Je remercie les personnes avec qui on partage tous les jours un repas dans la bonne

ambiance : Jean-Luc, Eric, Serge, Augustin, Gilles, Antoine, Heloise, Léa, Zouhir. En fait, je les remercie de me supporter et de supporter mes blagues surtout en période de solde avec le blague-Friday. Je remercie Shiyu, Pierre (tu n'as toujours pas fait ton pot de départ), Fabien et Matthieu. Merci à Cornilleau de partager son savoir aux jeunes. Il a su nous transmettre ses valeurs comme régler les problèmes autour d'une table. Nous avons pu passer un temps raisonnable à améliorer notre savoir-faire grâce à ses conseils avisés.

Je remercie mes amis Charlène, Seb et Paul. Merci à la célèbre paléobotaniste Anne-Laure de bien vouloir monter un projet de recherche avec moi sur les déjections de dinosaure. Je remercie Vincent d'avoir été en Thaïlande le jour de ma soutenance ! J'avoue avoir été en Corse pour la sienne... Je remercie Gilles le Breton, qui m'a toujours dit que "la thèse n'est pas une finalité en soi". Cette phrase m'a aidé tout au long de ma thèse. D'ailleurs, je n'ai jamais vraiment compris pourquoi la thèse devait être en soie alors que je l'ai imprimée sur du papier. Je remercie Audrey d'avoir partagé avec moi cette aventure. Je remercie également sa famille pour tout ce qui est logistique et pour les pizzas du vendredi. Je remercie Lionel, un fraté modèle ainsi que Chris et Raf. Je vous remercie vous, lecteur qui a parcouru ce long texte qui manque peut-être de formes $\perp$ et de couleurs.

Résumé

L'objectif de la thèse est d'explorer le potentiel d'un couplage entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale et forte résolution spatiale pour la sélection variétale. Ce système est étudié dans le cadre du phénotypage du maïs en conditions de stress hydrique. L'étude est organisée de la manière suivante :

1. Dans un premier temps, il s'agissait de vérifier l'hypothèse selon laquelle l'utilisation d'une forte résolution spectrale apporte un plus pour le phénotypage dans le cadre de la sélection variétale. Pour cela, deux campagnes expérimentales ont été réalisées en 2017 et 2018. Des spectres ont été acquis au champ en utilisant la spectroscopie visible et proche-infrarouge selon un plan d'expérience comptant au total 10 génotypes connus pour leur tolérance face au stress hydrique. Cette partie montre qu'il est possible de caractériser les comportements des génotypes en situation de stress hydrique tout en décrivant précisément les régions spectrales responsables de cette classification.
2. L'utilisation d'un spectromètre en extérieur induit un manque de répétabilité des mesures. Les conclusions des analyses réalisées sur des spectres portant cette erreur peuvent alors être faussées. Une méthode a donc été développée pour réduire l'erreur de répétabilité à travers l'utilisation d'une série de répétitions de mesures additionnelles au plan d'expérience. Cette méthode modifie l'algorithme d'analyse de variance ASCA en introduisant des projections orthogonales dans l'espace des spectres, en complément des projections orthogonales dans l'espace des individus, réalisées naturellement par l'analyse de variance.
3. L'objectif de la dernière partie était de réaliser le couplage d'un capteur à haute résolution spectrale et faible résolution spatiale (spectromètre Vis-NIR) avec un capteur à faible résolution spectrale et haute résolution spatiale (une caméra RGB) à l'aide d'algorithmes de pan-sharpening pour reconstituer une image hyperspectrale de test. Cette partie comportait deux étapes : une approche par simulation pour comparer les algorithmes de pan-sharpening à l'aide d'une image hyperspectrale et une partie pour proposer une méthode pour reconstruire une image hyperspectrale à partir de la solution de couplage proposée.

Abstract

The objective of the thesis is to explore the potential for coupling a high spectral resolution/low spatial resolution sensor and a low spectral resolution/high spatial resolution sensor for plant breeding. This system is studied in the context of maize breeding under water stress conditions. The study is organized as follows :

1. The first step was to test the hypothesis that the use of high spectral resolution is an advantage for phenotyping in varietal selection. For this purpose, two experimental campaigns were carried out in 2017 and 2018. Spectra were acquired in the field using visible and near-infrared spectroscopy according to an experimental design with a total of 10 genotypes known for their tolerance to water stress. This part shows how to describe genotypes behaviour under water stress conditions by accurately describing the spectral regions responsible for this classification.
2. The use of an outdoor spectrometer leads to a lack of repeatability of measurements. To perform analysis of variance of multivariate data with repeatability error can lead to wrong conclusions. A method has therefore been developed to reduce the repeatability error through the analysis of repeated measures of additional measurements to the experimental design. This method modifies the ASCA analysis of variance algorithm by introducing orthogonal projections into the row-space, in addition to the orthogonal projections into the column-space, which are naturally performed by analysis of variance.
3. The objective of the last part was to couple a high spectral resolution and low spatial resolution sensor (Vis-NIR spectrometer) with a low spectral resolution and high spatial resolution sensor (RGB camera) using pan-sharpening algorithms to reconstruct a hyperspectral test image. This part consisted of two steps : to compare pan-sharpening algorithms using a hyperspectral image by using simulation approach and to propose a method to reconstruct a hyperspectral image from the proposed coupling solution.

Table des matières

1 Introduction **13**

1.1 La sélection variétale 13

1.1.1 Contexte global : La sélection variétale en grandes cultures 13

1.1.2 L'enjeu du phénotypage 14

1.2 Contexte de la thèse 15

1.2.1 Objectifs de recherche et développement de Limagrain 15

1.2.2 Schéma de sélection du maïs 16

1.2.3 Recherche de nouveaux traits phénotypiques 17

1.2.4 Phénotypage par instrument optique 18

1.3 Objectifs opérationnels et questions scientifiques de la thèse 19

1.3.1 Objectifs opérationnels 19

1.3.2 Questions scientifiques 20

2 Pan-sharpening **25**

2.1 Introduction 26

2.2 Principes génériques et revue des différentes méthodes de pan-sharpening 28

2.2.1 Notations 28

2.2.2 Méthodes de substitution de composante (CS) 29

2.2.3 Méthodes par Analyse Multi-Résolution (MRA) 30

2.2.4 Méthodes basées sur le démélange spectral 31

2.3 Matériel et méthodes 34

2.3.1 Démarche générale 34

2.3.2 Indices utilisés pour l'évaluation de l'erreur de reconstruction . . . 34

2.3.3 Image hyperspectrale de référence 35

2.3.4 Cas d'étude 36

Influence de la résolution spatiale de l'image spectrale 36

Influence de la résolution spatiale de l'image panchromatique . . 37

2.4 Résultats et Discussions 37

2.4.1 Influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction et sur le temps de calcul 37
Qualité de reconstruction 38
2.4.2 Influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction et temps de calcul 43
2.4.3 Conclusion partielle 44
2.5 Problème de recouvrement 45
2.5.1 Méthode proposée pour le manque de recouvrement 45
Approche CNMF-KNN 46
Approche CNMF-PLS 46
2.5.2 Comparaison des deux méthodes proposées 47
Fusion d'une image spectrale avec une image panchromatique selon différents taux de recouvrement 47
Fusion d'une image spectrale avec une image multispectrale selon un taux de recouvrement défini 50
2.5.3 Conclusion partielle 53
2.6 Conclusion du chapitre 53

3 Proposition d'une méthode de correction des erreurs de répétabilité pour l'analyse de variance 61
3.1 Introduction 61
3.2 Prérequis et Définitions 62
3.2.1 Plan d'expérience et analyse de variance 62
3.2.2 Espace colonne et ligne ou Column and Row space 63
3.2.3 Orthogonalisation 64
3.3 Théorie de REP-ASCA 65
3.3.1 ASCA 65
3.3.2 Erreur de répétabilité 66
3.3.3 REP-ASCA 67
3.4 Test de REP-ASCA par simulation 68
3.4.1 Préparation de données simulées 69
3.4.2 Résultats et discussion 69
3.4.3 Conclusion de la simulation 71
3.5 Conclusion du chapitre 72

4 Potentiel d'une forte résolution spectrale pour la sélection variétale 77
4.1 Introduction 78
4.2 Matériels et méthodes 78

4.2.1 Dispositif expérimental 78
4.2.2 Données météorologiques et agronomiques collectées 79
4.2.3 Mesures spectrales 79
4.2.4 Protocole de mesures spectrales 81
4.3 Résultats et discussion 81
4.3.1 Description des variables agronomiques 82
4.3.2 Caractérisation des génotypes sur la base des données agronomiques 83
4.3.3 Caractérisation des conditions climatiques 84
4.3.4 Application de REP-ASCA 86
Terme traitement 89
Terme génotype 91
Terme interaction 93
4.4 Recherche de traits phénotypiques 96
4.4.1 Terme traitement 96
4.4.2 Terme génotype 97
4.4.3 Terme interaction 98
4.5 Conclusion du chapitre 100

5 Application des proxis obtenus : influence la date de mesure et de l'environnement étudié 107
5.1 Introduction 107
5.2 Influence de la date de mesure sur les proxis 109
5.2.1 Application sur les autres dates de mesures 109
Terme traitement 109
Terme génotype 110
Terme interaction 113
5.3 Influence de l'environnement sur les proxis 114
5.3.1 Description de la campagne expérimentale 2017 115
5.3.2 Application des composantes sur un autre environnement 116
Terme traitement 116
Terme génotype 117
Terme interaction 118
5.4 Conclusion du chapitre 120

6 Conclusion générale 125

A Annexes 129
A.1 Campagne expérimentale 2019 129

A.1.1 Introduction 129
A.1.2 Matériels et méthodes 130
Vol de drone 130
Traitement des données 130
A.1.3 Résultats et discussions 131
Description de la campagne expérimentale 2019 131
Résultats de REP-ASCA 132
A.1.4 Conclusion 133
A.2 Article sur la méthode REP-ASCA 134

B Communications **167**
B.1 Communications aux congrès 167
B.2 Articles publiés dans des revues scientifiques à comité de lecture 167

Table des figures **171**

Liste des tableaux **174**

Introduction générale

Créer de nouvelles variétés est un enjeu majeur pour l'agriculture pour faire face aux évolutions sociales et environnementales. La création de variétés passe par une étape de sélection qui consiste à croiser des variétés existantes et observer les phénotypes résultant de ce croisement directement au champ. L'étape d'observation du comportement des nouvelles variétés, appelée phénotypage, nécessite de parcourir de grandes surfaces cultivées. Cette étape représente alors un temps et des coûts élevés.

Selon les objectifs de sélection, un suivi temporel des variétés peut être mis en place. Il y a donc une tendance actuelle à utiliser des capteurs non destructifs pour observer ces phénotypes. Ces dernières années, le phénotypage par instrument optique s'est développé et plus particulièrement avec l'imagerie multispectrale. Lorsque cette solution est associée au vecteur drone, il répond aux contraintes de phénotypage haut-débit pour la sélection variétale. En effet, cela permet de parcourir de grandes surfaces très rapidement. L'image obtenue fournit une information spatiale qui peut être exploitée. Des études se basent alors sur l'information spatiale générée comme la détection de forme ou de texture. Pour des raisons technologiques et financières, ces capteurs ont une faible résolution spectrale. Les caméras multispectrales fournissent alors quelques bandes spectrales pouvant être combinées pour calculer des indices de végétation. En ne retenant quelques bandes spectrales, on se prive d'information contenue dans le spectre de réflectance. L'imagerie dite hyperspectrale (forte résolution spectrale et spatiale) fournit énormément d'information spatiale et spectrale. Cette technologie ne peut pas être associée au drone pour des raisons de coûts et de problèmes liés à l'acquisition (temps d'acquisition, poids, ...).

Il est courant, à partir d'une technologie de forte résolution spectrale, de ne retenir que les bandes spectrales à des longueurs d'onde spécifiques. Nous avons fait le choix dans cette thèse, d'utiliser les spectres directement comme des critères de phénotypage pouvant dissocier les génotypes ou les caractériser. En ne dégradant pas l'information spectrale, on ne rend pas spécifique le capteur.

La solution proposée dans la thèse est de coupler deux capteurs pour obtenir une solution d'imagerie comparable à de l'imagerie hyperspectrale. Le pari relevé par cette thèse est double : premièrement, prouver qu'une forte résolution spectrale peut être

utile pour le phénotypage. Et deuxièmement, concevoir un imageur léger et peu coûteux comparable à de l'imagerie hyperspectrale.

Le manuscrit est organisé en 6 chapitres. Le premier est l'introduction qui explique, dans un premier temps, le contexte et l'enjeu du phénotypage dans la sélection variétale, puis, dans un second temps, les attentes opérationnelles et les questions scientifiques reliées à la thèse. Le deuxième chapitre appelé 'pan-sharpening' traite spécifiquement de la création d'une image de type hyperspectrale. Pour cela, des algorithmes issus de la télédétection ont été adaptés à un dispositif constitué d'une caméra couleur et d'un spectromètre afin de fournir une image avec une forte résolution spectrale et une forte résolution spatiale. Le troisième chapitre propose une méthode d'analyse de variance, basée sur une méthode existante (ASCA) et adaptée aux mesures spectrales de végétation en extérieur. La méthode développée dans ce chapitre propose de tenir compte de l'erreur de répétabilité de manière explicite, afin d'améliorer les performances de la séparation des sources de variances portées par un plan d'expériences. Le quatrième chapitre étudie le potentiel d'une forte résolution spectrale pour la sélection variétale. Un prototype de mesure de réflectance a été utilisé pour mesurer les spectres de chaque micro-parcelle de maïs dans un dispositif expérimental croisant 10 génotypes avec 2 niveaux de contraintes hydriques. Les données recueillies ont été analysées avec la méthode proposée en troisième chapitre, mettant en évidence les signatures spectrales reliées à l'interaction génotype $\times$ environnement. Le cinquième chapitre reporte le résultat de l'application du modèle du chapitre 4 sur des données acquises dans un autre lieu et une autre année confirmant l'applicabilité de la méthode développée. Le dernier chapitre apporte les conclusions générales et discute des résultats.

Introduction

Chapitre 1

Sommaire

1.1 La sélection variétale 13

1.1.1 Contexte global : La sélection variétale en grandes cultures . 13

1.1.2 L'enjeu du phénotypage 14

1.2 Contexte de la thèse 15

1.2.1 Objectifs de recherche et développement de Limagrain . . . 15

1.2.2 Schéma de sélection du maïs 16

1.2.3 Recherche de nouveaux traits phénotypiques 17

1.2.4 Phénotypage par instrument optique 18

1.3 Objectifs opérationnels et questions scientifiques de la thèse 19

1.3.1 Objectifs opérationnels 19

1.3.2 Questions scientifiques 20

1.1 La sélection variétale

1.1.1 Contexte global : La sélection variétale en grandes cultures

Dans les années à venir, une élévation de la température globale est attendue entraînant une augmentation de la demande évaporative de l'air. Les épisodes de sécheresse seront de plus en plus sévères et de plus en plus fréquents entraînant un affaiblissement des réserves hydriques dans le sol. Cela réduira considérablement le rendement agricole (Lobell et al., 2011; Hawkins et al., 2012) dans une période croissante de demande alimentaire mondiale (Tilman et al., 2011).

Les défis scientifiques du monde agricole sont de proposer de nouvelles solutions s'inscrivant dans un contexte d'adaptation au changement climatique. Des améliorations sont

attendues pour l'ensemble des formes d'agriculture (agriculture biologique, conventionnelle, biodynamique ...). L'adaptation au changement climatique est donc une priorité pour l'agriculture et notamment pour les grandes cultures qui doivent assurer la sécurité alimentaire mondiale.

C'est dans ce contexte que la sélection variétale doit s'adapter et permettre la création de variétés présentant une combinaison optimale des caractères phénotypiques (physiologiques ou morphologiques) tout en étant en adéquation avec son environnement. Aujourd'hui les critères de sélection se sont diversifiés comme des critères de stabilité du rendement face à des conditions climatiques non prédictibles ou encore des critères de tolérance face aux maladies ou à des stress abiotiques tout en associant performance économique et performance environnementale (Van Bueren et al., 2002).

1.1.2 L'enjeu du phénotypage

Au cours de ces dernières années, la sélection variétale a évolué en lien avec les progrès réalisés en biologie moléculaire. Accéder à une grande information génétique a un coût de plus en plus faible (Elshire et al., 2011). A la différence du génotype qui contient l'information portée par le génome, le phénotype est défini comme les traits observables. Il se caractérise à la fois par l'expression du code génétique mais aussi de l'interaction avec des facteurs environnementaux. La collection d'informations phénotypiques (appelée phénotypage) possède de nombreuses contraintes opérationnelles liées au nombre important de génotypes testés dans un même essai (plusieurs milliers).

FIGURE 1.1 – Exemples de micro-parcelles.

En effet, chaque génotype sera testé dans un dispositif de petites parcelles expérimentales appelées aussi micro-parcelles (Fig. 1.1) semées sur quelques rangées (de 2 à 4) de quelques mètres de long (entre 6 et 12m). Les observations doivent être assez précises pour caractériser une micro-parcelle et couvrir l'ensemble du lieu d'essai. Seulement obtenir des détails sur ces grandes surfaces est coûteux et chronophage. Le phénotypage est aujourd'hui le processus limitant de la sélection variétale (Araus and

Cairns, 2014; Tuberosa, 2012). Pour répondre aux contraintes de coût et de haut-débit, il est nécessaire de développer des outils précis, rapides et agiles.

Au cours des dernières années, des plateformes de phénotypage haut-débit dans des milieux contrôlés ont vu le jour : PHENOARCH (Cabrera Bosquet et al., 2015), PHENOPSIS (Granier et al., 2006), plateforme de phénotypage racine (Clark et al., 2013), TraitMill (Reuzeau et al., 2010), SPICY (van der Heijden et al., 2012). L'acquisition des données phénotypiques en milieux contrôlés montre plusieurs avantages comme le paramétrage précis des conditions de stress ou l'étude spécifique de plusieurs génotypes dans les mêmes conditions. Il est facile d'automatiser les acquisitions tout au long du cycle de vie d'une plante. Cependant, dans la plupart des cas, ces conditions ne sont pas totalement représentatives d'une parcelle agricole (Araus et al., 2008; White et al., 2012) et il est difficile d'extrapoler le comportement au champ à partir d'observations réalisées en conditions contrôlées (Araus and Cairns, 2014; Passioura, 2006).

1.2 Contexte de la thèse

1.2.1 Objectifs de recherche et développement de Limagrain

Limagrain est un groupe coopératif agricole international employant plus de 10000 salariés dans le monde répartis au sein de 56 pays. Le groupe fondé en Auvergne en 1965 et est basé à Saint-Beauzire dans le Puy-de-Dôme. Il représente actuellement le 4ème semencier mondial avec un chiffre d'affaires de 2.6 milliards d'euros dont 14.6% est investi dans la recherche. Les activités du groupe sont réparties entre les semences de grandes cultures (maïs, blé, colza, tournesol...), les semences potagères et les produits céréaliers (ingrédients céréaliers et boulangerie-pâtisserie). Les activités de semences de grandes cultures en Europe sont rassemblées au sein de Limagrain Europe, filiale du groupe Limagrain.

L'équipe Native Traits (NT) est une équipe de recherche et développement multi-espèces tournée vers deux thématiques : le phénotypage haut-débit et l'envirotypage [1]. L'objectif de l'équipe est de proposer de nouvelles méthodes et de nouveaux outils aux sélectionneurs pour la création de variétés améliorées pour leurs tolérances aux stress abiotiques.

1. L'envirotypage est la caractérisation des conditions environnementales susceptibles d'agir sur une espèce

1.2.2 Schéma de sélection du maïs

Après avoir identifié clairement les objectifs de sélection (rendement, maturité, stabilité, adaptation à des marchés spécifiques, tolérance, maladie ...), les sélectionneurs de Limagrain suivent un schéma général de sélection et d'expérimentation pour développer de nouvelles lignées, parents des futurs hybrides.

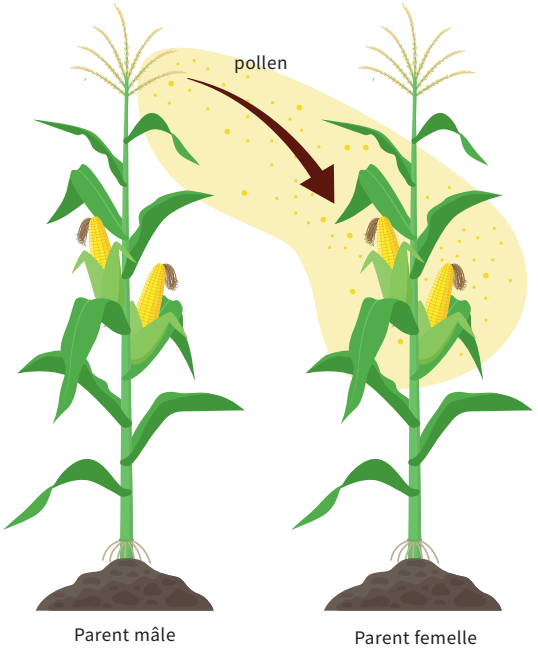


FIGURE 1.2 – Fécondation croisée entre parents mâle et femelle.

Aujourd'hui, les semences de maïs commercialisées sont des maïs dits hybrides. Ces hybrides sont obtenus alors par fécondation croisée entre deux lignées parentales [2] (Fig. 1.2). Les hybrides, alors hétérozygotes, sont bien plus productifs et vigoureux que les lignées, c'est ce que l'on appelle *l'hétérosis* (East, 1908; Shull, 1909).

Les hybrides sont commercialisés sur l'ensemble de l'Europe, mais se distinguent du Nord au Sud par leur précocité, maïs dit précoce au Nord,et tardif au Sud. Plus un maïs est précoce, moins il aura besoin de jours pour achever son cycle de développement, aux dépens du potentiel de rendement. Au-delà de la maturité, des zones de marchés sont définies selon des cibles de sélection spécifiques. Les hybrides sont alors évalués sur un réseau commun d'essais au champ représentatif de chacune de ces zones de marché. Ainsi les hybrides destinés à des zones limitantes en eau sont testés sur des réseaux intégrant à la fois des lieux stressés et des lieux à potentiel.

2. Les lignées parentales appelées également lignées pures sont les souches génétiques. Cela signifie que les individus d'une lignée pure sont génétiquement identiques.

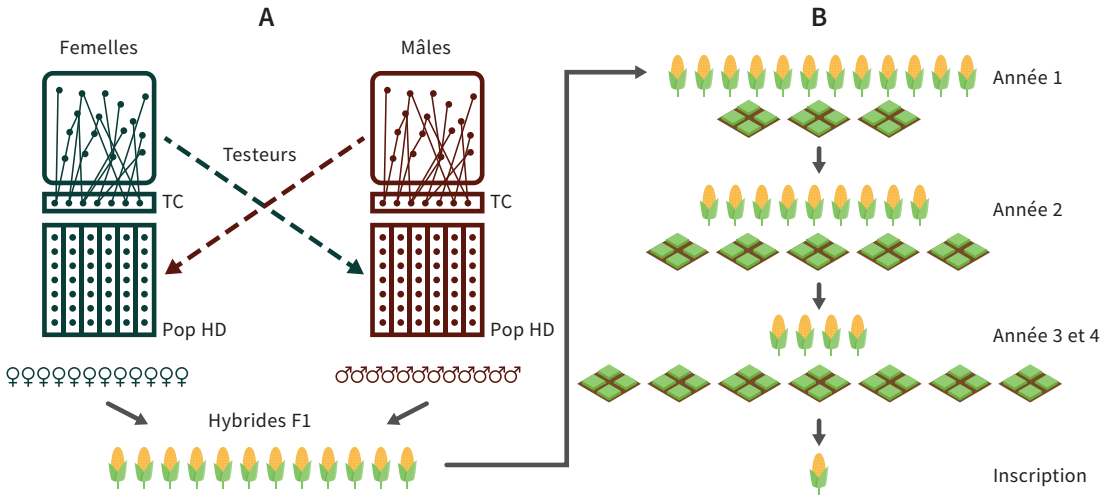


FIGURE 1.3 – Schéma général de sélection du maïs.

Le schéma de sélection du maïs se décompose en deux étapes (Fig. 1.3). La première étape consiste à sélectionner les lignées parentales afin de combiner leurs caractéristiques propres. Les lignées mâles et femelles sont sélectionnées en parallèle dans des programmes différents selon les traits spécifiques par groupe hétérotique. On évalue leur Aptitude Générale à la Combinaison (AGC) en les croisant avec des génotypes représentatifs du groupe sexuel opposé, appelés testeurs. L'AGC correspond à la valeur de rendement moyenne de la descendance de la lignée avec l'ensemble des testeurs.

La deuxième étape consiste à évaluer les hybrides sur plusieurs lieux et pendant quatre années pour vérifier leur stabilité et leur adaptation agronomique aux segments de marchés ciblés. Cette étape consiste à évaluer de nombreuses combinaisons hybrides issues de la combinaison des lignées parentales qui ont les meilleures valeurs d'AGC. A chaque génération, seulement 10% des hybrides sont sélectionnés sur la base de critères de rendement et d'adaptation agronomique de plus en plus précis, pour être testés l'année suivante. En parallèle à ce processus de sélection, les futurs hybrides commerciaux vont être testés sur un nombre croissant de lieux d'essais.

Les deux dernières années correspondent à l'évaluation technique pour éventuellement inscrire et commercialiser 1 ou 2 génotypes apportant une amélioration par rapport aux variétés déjà présentes sur le marché.

1.2.3 Recherche de nouveaux traits phénotypiques

Une fois l'objectif défini, le sélectionneur établit un protocole visant à mesurer un ensemble de traits agronomiques. De nos jours, les sélectionneurs ont accès à certains

nombres d'outils permettant de mesurer de la façon la plus objective possible des traits caractéristiques de chaque variété. De nombreux traits sont autant d'information permettant de sélectionner efficacement les nouvelles variétés.

Sur l'ensemble des traits agronomiques existants, le rendement et l'humidité sont obtenus au moment de la récolte tandis que d'autres traits nécessitent un suivi avec des mesures effectuées tout au long du cycle de la plante : stay green, dénombrement des plantes dans la parcelle,notations qualitatives des parcelles (vigueur départ, qualité d'émergence ...). Ce processus d'acquisition de données phénotypiques est long et coûteux. Du fait de la nature manuelle de certaines de ces mesures, les données collectées sont sujettes aux erreurs humaines, au manque de répétabilité ou au biais induit par l'expérimentateur.

Des solutions telles que les services de mesures optiques par télédétection satellite permettent d'accéder à des informations sur la parcelle mais sont inadaptées pour de la sélection. En effet, des difficultés liées à l'imagerie satellitaire (correction atmosphérique, période de revisite trop élevée, présence de nuages...) et la très faible résolution spatiale ne permettent pas un suivi temporel de micro-parcelles. Des vecteurs mobiles aéroportés permettent d'éviter ces problèmes. Le ballon dirigeable représente la solution la moins onéreuse, mais n'est pas réellement adapté à la sélection variétale pour des raisons opérationnelles (réglementation, encombrement, déplacements non précis, instabilité, charge limite ...). L'utilisation d'aéronefs a l'avantage de pouvoir embarquer des caméras très lourdes et de réaliser des acquisitions sur de grandes surfaces. Néanmoins le coût reste trop élevé lorsqu'il s'agit de réaliser plusieurs vols (Deery et al., 2014).

Le vecteur drone est considéré depuis quelques années comme un outil idéal de phénotypage haut-débit (von Bueren et al., 2015). Des capteurs embarqués sur drone permettent ainsi d'accéder rapidement aux traits caractéristiques (Sankaran et al., 2015). Parmi les capteurs utilisés, on peut recenser les capteurs thermiques, optiques et le lidar.

1.2.4 Phénotypage par instrument optique

Entre deux génotypes, la différence de réponse aux conditions environnementales peut être infime. L'outil de phénotypage doit être assez performant pour caractériser cette différence tout en étant assez fiable pour référencer les positions de chaque micro-parcelle.

Les capteurs multispectraux légers embarqués sur drone permettent de cartographier de manière précise les parcelles d'expérimentation. Des indices appelés indices de végétation sont calculés à partir de ces images en combinant plusieurs bandes spectrales. Les plus connus sont le Normalized Differential Vegetation Index (NDVI) et le Photochemical Reflectance Index (PRI) Ces indices sont obtenus à partir d'une combinaison

de deux bandes spectrales et sont connus pour être corrélés à la biomasse, la teneur en chlorophylle ou à l'activité photosynthétique (Gamon et al., 1995; Cabrera-Bosquet et al., 2011). L'imagerie multispectrale sur drone permet ainsi le suivi de ces indices au cours des différents stades phénologiques (Baret and Guyot, 1986; Labus et al., 2002). L'analyse de l'évolution de ces indices permet d'accéder à des traits tels que l'entrée en sénescence, le stay green ou la vigueur départ.

La dégradation de la résolution spectrale [3] est la solution la plus utilisée, car elle permet de fabriquer des capteurs légers et peu coûteux. Cependant, en réduisant fortement la résolution spectrale, on se prive d'une partie de l'information spectrale contenue dans le spectre de réflectance. Le calcul d'indice peut alors mener à une interprétation erronée ou limitée du comportement du matériel végétal (Römer et al., 2012).

L'imagerie hyperspectrale représente la meilleure solution pour accéder à une information riche à la fois spatialement et spectralement (Vigneau et al., 2011). L'utilisation d'une forte résolution spectrale permet d'apporter plus d'information. Des études ont montré que l'utilisation du spectre apportait une plus-value aux indices de végétation pour la prédiction d'un ensemble de traits (chlorophylle, azote, saccharose, pression saturante ...) (Yendrek et al., 2016; Hansen and Schjoerring, 2003; Blackburn and Ferwerda, 2008) ou la détection précoce de maladie (Behmann et al., 2014). L'hypothèse formulée à travers la littérature est que l'information contenue dans le spectre permet d'accéder à des détails plus précis que l'utilisation d'indices de végétation dans le cadre de la sélection variétale. Néanmoins, l'imagerie hyperspectrale souffre de nombreux désavantages pour des raisons opérationnelles : poids de la caméra, mesure de la référence, temps d'acquisition et coût. Pour ces raisons, il est difficile d'envisager l'utilisation de ce type de caméra sur drone dans le cadre de phénotypage pour de la sélection variétale.

Le verrou technologique lié à l'incapacité de combiner drone et caméra hyperspectrale pour la sélection variétale définit ainsi le cadre de cette étude scientifique.

1.3 Objectifs opérationnels et questions scientifiques de la thèse

1.3.1 Objectifs opérationnels

Accéder à une information spectrale plus riche à travers des solutions peu coûteuses telles que la spectroscopie Infrarouge (IR) doit permettre de différencier les génotypes de manière plus précise. L'objectif de la thèse est d'explorer le potentiel d'un couplage

3. La dégradation spectrale est définie comme l'utilisation de quelques bandes spectrales à partir d'un spectre continu.

entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale et forte résolution spatiale afin de reconstituer une image hyperspectrale pour du phénotypage haut-débit au champ en sélection variétale.

1.3.2 Questions scientifiques

Dans cette étude, on fait l'hypothèse qu'une information à forte résolution spectrale est un plus pour le phénotypage en sélection variétale. L'objectif de la thèse est alors de répondre aux questions scientifiques suivantes :

— Comment peut-on coupler deux capteurs pour obtenir une information aussi riche que de l'imagerie hyperspectrale ?

— Comment traiter les plans d'expériences de phénotypage quand la grandeur acquise est un spectre de réflectance acquis en extérieur ?

— Comment obtenir des traits phénotypiques en tenant compte de la forte résolution spectrale ?

La première question propose d'anticiper la solution de couplage de capteurs indépendamment de l'application visée pour évaluer la faisabilité technologique et le potentiel du système proposé. Les deux questions suivantes traitent spécifiquement de l'application d'une telle solution pour du phénotypage dans le cadre de la sélection variétale.

Dans un premier temps, nous allons étudier comment reconstruire une image hyperspectrale à partir de la solution de couplage étudiée : à savoir, un capteur imageur RGB et un spectromètre. Cette étape permettra de vérifier la faisabilité opérationnelle du couplage proposé. Pour cela, nous étudierons les algorithmes de fusion d'images appelés méthode de pan-sharpening. Dans un second temps, nous proposerons une méthode de type analyse de variance pour réduire des erreurs de mesures liées au problème de répétabilité des mesures. Enfin, nous vérifierons qu'une forte résolution spectrale permet de mieux discriminer les génotypes soumis à un stress hydrique. Pour cela, nous appliquerons la méthode proposée à un jeu de données issues de deux campagnes expérimentales réalisées en 2017 et 2018. Nous obtiendrons ainsi des signatures spectrales associées à des variables d'influences (traitement, génotype, etc.). Ces signatures spectrales seront utilisées pour classer les variétés et identifier parmi elles les plus tolérantes.

Bibliographie

David B. Lobell, Wolfram Schlenker, and Justin Costa-Roberts. Climate trends and global crop production since 1980. *Science*, 333(6042) :616–620, 2011. URL http://science.sciencemag.org/content/333/6042/616.short.

Ed Hawkins, Thomas E. Fricker, Andrew J. Challinor, Christopher AT Ferro, Chun Kit Ho, and Tom M. Osborne. Increasing influence of heat stress on French maize yields from the 1960s to the 2030s. 2012.

D. Tilman, C. Balzer, J. Hill, and B. L. Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, 108 (50) :20260–20264, December 2011. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1116437108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1116437108.

ET Lammerts Van Bueren, P. C. Struik, and E. Jacobsen. Ecological concepts in organic farming and their consequences for an organic crop ideotype. *NJAS-Wageningen Journal of Life Sciences*, 50(1) :1–26, 2002. URL http://www.sciencedirect.com/science/article/pii/S157352140280001X.

Robert J. Elshire, Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S. Buckler, and Sharon E. Mitchell. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5) :e19379, 2011.

José Luis Araus and Jill E. Cairns. Field high-throughput phenotyping : the new crop breeding frontier. *Trends in Plant Science*, 19(1) :52–61, January 2014. ISSN 13601385. doi : 10.1016/j.tplants.2013.09.008. URL http://linkinghub.elsevier.com/retrieve/pii/S1360138513001994.

Roberto Tuberosa. Phenotyping for drought tolerance of crops in the genomics era. 2012.

Llorenç Cabrera Bosquet, Nicolas Brichet, Christian Fournier, Antonin Grau, Jonathan Mineau, Vincent Negre, Pascal Neveu, Benoit Suard, Anne Tireau, Claude Welcker, and Francois Tardieu. PHENOARCH, a multiscale phenotyping platform for plant architecture, growth rate, water use efficiency and radiation use efficiency. *2015 ; Recent progress in drought tolerance from genetics to modelling, Montpellier, FRA, 2015-06-08-2015-06-09, 81*, 2015. URL http://agris.fao.org/agris-search/search.do?recordID=LV2016028780.

Christine Granier, Luis Aguirrezabal, Karine Chenu, Sarah Jane Cookson, Myriam Dauzat, Philippe Hamard, Jean-Jacques Thioux, Gaëlle Rolland, Sandrine Bouchier-Combaud, and Anne Lebaudy. PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. *New phytologist*, 169(3) :623–635, 2006.

Randy T. Clark, Adam N. Famoso, Keyan Zhao, Jon E. Shaff, Eric J. Craft, Carlos D. Bustamante, Susan R. Mccouch, Daniel J. Aneshansley, and Leon V. Kochian. High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development : Root phenotyping platform. *Plant, Cell & Environment*, 36(2) :454–466, February 2013. ISSN 01407791. doi : 10.1111/j.1365-3040.2012.02587.x. URL http://doi.wiley.com/10.1111/j.1365-3040.2012.02587.x.

Christophe Reuzeau, Jan Pen, Valerie Frankard, Joris de Wolf, Rindert Peerbolte, Willem Broekaert, and Wim van Camp. TraitMill : a discovery engine for identifying yield-enhancement genes in cereals. *Plant Gene and Trait*, 1, 2010.

Gerie van der Heijden, Yu Song, Graham Horgan, Gerrit Polder, Anja Dieleman, Marco Bink, Alain Palloix, Fred van Eeuwijk, and Chris Glasbey. SPICY : towards automated phenotyping of large pepper plants in the greenhouse. *Functional Plant Biology*, 39 (11) :870, 2012. ISSN 1445-4408. doi : 10.1071/FP12019. URL http://www.publish.csiro.au/?paper=FP12019.

José Luis Araus, Gustavo A. Slafer, Conxita Royo, and M. Dolores Serret. Breeding for Yield Potential and Stress Adaptation in Cereals. *Critical Reviews in Plant Sciences*, 27(6) :377–412, November 2008. ISSN 0735-2689, 1549-7836. doi : 10.1080/07352680802467736. URL http://www.tandfonline.com/doi/abs/10.1080/07352680802467736.

Jeffrey W. White, Pedro Andrade-Sanchez, Michael A. Gore, Kevin F. Bronson, Terry A. Coffelt, Matthew M. Conley, Kenneth A. Feldmann, Andrew N. French, John T. Heun, Douglas J. Hunsaker, Matthew A. Jenks, Bruce A. Kimball, Robert L. Roth, Robert J. Strand, Kelly R. Thorp, Gerard W. Wall, and Guangyao Wang. Field-based phenomics for plant genetics research. *Field Crops Research*, 133 :101–112, July 2012. ISSN 03784290. doi : 10.1016/j.fcr.2012.04.003. URL http://linkinghub.elsevier.com/retrieve/pii/S037842901200130X.

John B. Passioura. The perils of pot experiments. *Functional Plant Biology*, 33(12) : 1075–1079, 2006.

Edward M. East. Inbreeding in corn. *Rep. Conn. Agric. Exp. Stn*, 1907 :419–428, 1908.

George Harrison Shull. A pure-line method in corn breeding. *Journal of Heredity*, (1) : 51–58, 1909.

David Deery, Jose Jimenez-Berni, Hamlyn Jones, Xavier Sirault, and Robert Furbank. Proximal Remote Sensing Buggies and Potential Applications for Field-Based Phenotyping. *Agronomy*, 4(3) :349–379, July 2014. ISSN 2073-4395. doi : 10.3390/agronomy4030349. URL http://www.mdpi.com/2073-4395/4/3/349/.

S. K. von Bueren, A. Burkart, A. Hueni, U. Rascher, M. P. Tuohy, and I. J. Yule. Deploying four optical UAV-based sensors over grassland : challenges and limitations. *Biogeosciences*, 12(1) :163–175, January 2015. ISSN 1726-4189. doi : 10.5194/bg-12-163-2015. URL http://www.biogeosciences.net/12/163/2015/.

Sindhuja Sankaran, Lav R. Khot, Carlos Zúñiga Espinoza, Sanaz Jarolmasjed, Vidyasagar R. Sathuvalli, George J. Vandemark, Phillip N. Miklas, Arron H. Carter, Michael O. Pumphrey, N. Richard Knowles, and Mark J. Pavek. Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping : A review. *European Journal of Agronomy*, 70 :112–123, October 2015. ISSN 11610301. doi : 10.1016/j.eja.2015.07.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1161030115300095.

John A. Gamon, Christopher B. Field, Michael L. Goulden, Kevin L. Griffin, Anne E. Hartley, Geeske Joel, Josep Penuelas, and Riccardo Valentini. Relationships Between NDVI, Canopy Structure, and Photosynthesis in Three Californian Vegetation Types. *Ecological Applications*, 5(1) :28–41, February 1995. ISSN 10510761. doi : 10.2307/1942049. URL http://doi.wiley.com/10.2307/1942049.

L. Cabrera-Bosquet, G. Molero, A. Stellacci, J. Bort, S. Nogués, and J. Araus. NDVI as a potential tool for predicting biomass, plant nitrogen content and growth in wheat genotypes subjected to different water and nitrogen conditions. *Cereal Research Communications*, 39(1) :147–159, March 2011. ISSN 0133-3720, 1788-9170. doi : 10.1556/CRC.39.2011.1.15. URL http://www.akademiai.com/doi/abs/10.1556/CRC.39.2011.1.15.

Frederic Baret and Gérard Guyot. Suivi de la maturation de couverts de blé par radiométrie dans les domaines visible et proche infra-rouge. *Agronomie*, 6(6) :509–516, 1986. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=8753247.

M. P. Labus, G. A. Nielsen, R. L. Lawrence, R. Engel, and D. S. Long. Wheat yield estimates using multi-temporal NDVI satellite imagery. *International Journal of Remote Sensing*, 23(20) :4169–4180, January 2002. ISSN 0143-1161, 1366-5901. doi : 10.1080/01431160110107653. URL https://www.tandfonline.com/doi/full/10.1080/01431160110107653.

Christoph Römer, Mirwaes Wahabzada, Agim Ballvora, Francisco Pinto, Micol Rossini, Cinzia Panigada, Jan Behmann, Jens Léon, Christian Thurau, Christian Bauckhage, and others. Early drought stress detection in cereals : simplex volume maximisation for hyperspectral image analysis. *Functional Plant Biology*, 39(11) :878–890, 2012. URL http://www.publish.csiro.au/?paper=FP12060.

Nathalie Vigneau, Martin Ecarnot, Gilles Rabatel, and Pierre Roumet. Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat. *Field Crops Research*, 122(1) :25–31, April 2011. ISSN 03784290. doi : 10.1016/j.fcr.2011.02.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0378429011000451.

Craig Yendrek, Tiago Tomaz, Christopher M Montes, Youyuan Cao, Alison M Morse, Patrick J Brown, Lauren McIntyre, Andrew Leakey, and Elizabeth Ainsworth. High-throughput phenotyping of maize leaf physiology and biochemistry using hyperspectral reflectance. *Plant Physiology*, page pp.01447.2016, November 2016. ISSN 0032-0889, 1532-2548. doi : 10.1104/pp.16.01447. URL http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.01447.

P. M. Hansen and J. K. Schjoerring. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote sensing of environment*, 86(4) :542–553, 2003. URL http://www.sciencedirect.com/science/article/pii/S0034425703001317.

G Blackburn and J Ferwerda. Retrieval of chlorophyll concentration from leaf reflectance spectra using wavelet analysis. *Remote Sensing of Environment*, 112(4) : 1614–1632, April 2008. ISSN 00344257. doi : 10.1016/j.rse.2007.08.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0034425707004117.

Jan Behmann, Jörg Steinrücken, and Lutz Plümer. Detection of early plant stress responses in hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93 :98–111, 2014. URL http://www.sciencedirect.com/science/article/pii/S092427161400094X.

Pan-sharpening

Chapitre 2

Sommaire

2.1 Introduction **26**

2.2 Principes génériques et revue des différentes méthodes de pan-sharpening **28**

2.2.1 Notations 28

2.2.2 Méthodes de substitution de composante (CS) 29

2.2.3 Méthodes par Analyse Multi-Résolution (MRA) 30

2.2.4 Méthodes basées sur le démélange spectral 31

2.3 Matériel et méthodes **34**

2.3.1 Démarche générale 34

2.3.2 Indices utilisés pour l'évaluation de l'erreur de reconstruction 34

2.3.3 Image hyperspectrale de référence 35

2.3.4 Cas d'étude 36

Influence de la résolution spatiale de l'image spectrale . . 36

Influence de la résolution spatiale de l'image panchromatique 37

2.4 Résultats et Discussions **37**

2.4.1 Influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction et sur le temps de calcul 37

Qualité de reconstruction 38

Indices moyennés 38

Détails des indices SAM et CC pour les meilleures méthodes de chaque classe 40

Valeurs de l'indice SAM pixel à pixel 40

Valeurs de l'indice CC par longueur d'onde 41

Temps de calcul des différentes méthodes de pan-sharpening testées en fonction de N . . 42

2.4.2 Influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction et temps de calcul 43

Indices moyennés 43

Temps de calcul des différentes méthodes de pan-sharpening testées en fonction de M . . 43

2.4.3 Conclusion partielle 44
2.5 Problème de recouvrement 45
2.5.1 Méthode proposée pour le manque de recouvrement 45
Approche CNMF-KNN 46
Approche CNMF-PLS 46
2.5.2 Comparaison des deux méthodes proposées 47
Fusion d'une image spectrale avec une image panchromatique selon différents taux de recouvrement . . . 47
Fusion d'une image spectrale avec une image multispectrale selon un taux de recouvrement défini . . . 50
2.5.3 Conclusion partielle 53
2.6 Conclusion du chapitre 53

2.1 Introduction

Le **pan-sharpening** est l'augmentation de la résolution spatiale d'une image spectrale[1] à partir d'une image plus résolue spatialement. Le terme pan-sharpening vient de l'utilisation des informations spatiales contenues dans une image appelée image panchromatique ne contenant qu'une seule bande spectrale. Les méthodes dites de pan-sharpening font donc référence plus généralement aux méthodes de fusion de données issues de différents capteurs.

Historiquement, le pan-sharpening est apparu au sein de la communauté de la télédétection avec l'avènement des images satellites en provenance des satellites SPOT et Landsat (Cliche, 1985; Welch, 1985; Chavez Jr, 1984). Des méthodes ont donc été développées pour fusionner des données multispectrales avec des images panchromatiques (Vivone et al., 2015). Accéder à une information spatiale plus précise de données spectrales est utile dans de nombreuses applications comme l'identification de structures archéologiques (Lasaponara and Masini, 2012), la gestion durable de forêts (Ozdemir, 2008) ou le suivi des cultures (Rahaman et al., 2017).

Les images satellites possèdent généralement quelques bandes spectrales choisies pour divers domaines d'applications notamment dans le domaine agricole. La combinaison de bandes spectrales permet de reconstruire des indices appelés indices de végétation. Plusieurs indices corrélés à une teneur spécifique ou à des paramètres agronomiques ont été proposés dans la littérature. Lors d'acquisitions satellitaires, ces indices de végétation permettent de fournir une information spectrale à l'échelle d'une région (Labus et al., 2002), mais aussi à l'échelle de la parcelle (Colombo et al., 2003).

1. Une image spectrale est définie comme étant une image contenant plusieurs bandes spectrales.

Le pan-sharpening est une solution permettant d'augmenter la résolution spatiale et ainsi accéder à des informations plus précises dans la parcelle. Les méthodes de pan-sharpening sont généralement utilisées pour fusionner deux images provenant d'un satellite. Néanmoins, certaines études dans le domaine de l'agriculture de précision s'appuient sur l'utilisation des méthodes de pan-sharpening pour fusionner des images satellites avec des images issues de drones (Gevaert et al., 2015; Jenerowicz and Woroszkiewicz, 2016). En effet, l'utilisation d'un capteur sur drone fournit une information plus résolue spatialement que les données satellites. Cela permet par la suite d'accéder à des traits physiologiques et agronomiques à l'échelle de micro-parcelles expérimentales.

Toutefois, fusionner des données issues de drones avec un satellite peut entraîner certaines complications comme la latence d'acquisition entre les deux images ou des problèmes liés à la correction atmosphérique. De plus, il est difficile d'établir une correspondance entre les pixels [2] des deux capteurs. L'ensemble de ces problèmes peut entraîner des erreurs lors de la fusion des données (Blanc et al., 1998). Proposer la solution d'embarquer deux capteurs sur drone permet de respecter ces conditions pour appliquer des méthodes de pan-sharpening.

Dans un premier temps, nous décrirons dans ce chapitre un ensemble de méthodes de pan-sharpening parmi les plus utilisées. Puis, nous étudierons par simulation le potentiel d'un couplage de capteurs de résolutions spatiale et spectrale différentes. Cette démarche nous permettra de sélectionner la méthode de pan-sharpening la plus adaptée et d'anticiper des questions d'ordres opérationnelles :

1. Combien de spectres sont nécessaires pour reconstruire une image ?
2. Quelle est l'influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction ?
3. Comment reconstruire une image lorsque des pixels de l'image panchromatique ne sont pas associés à des mesures spectrales ?

Les trois questions dépendent des capteurs utilisés et de l'application visée. En effet, elles permettent d'anticiper la conception d'un outil et du protocole relié à l'acquisition finale. La dernière question traite spécifiquement le cas où le capteur portant une forte résolution spectrale possède une très faible résolution spatiale. En effet, il est possible que des pixels de l'image panchromatique ne soient pas associés à une mesure spectrale. Dans ce cas, nous proposons deux méthodes permettant d'attribuer des spectres à ces pixels.

2. appelée aussi appariement

2.2 Principes génériques et revue des différentes méthodes de pan-sharpening

Les premiers algorithmes de pan-sharpening ont été créés pour augmenter la résolution spatiale d'une image multispectrale avec peu de bandes spectrales. Parmi ces algorithmes, on peut distinguer deux classes de méthodes (Vivone et al., 2015) : la substitution de composante (CS) et l'analyse multi-résolution (MRA). Bien que développées pour le cas multispectral, ces approches sont également adaptées pour des images hyperspectrales (Loncan et al., 2015) [3].

Néanmoins, d'autres méthodes ont été conçues spécialement pour augmenter la résolution spatiale d'images hyperspectrales à partir d'image multispectrale ou panchromatique (Yokoya et al., 2012; Simões et al., 2015; Wei et al., 2015a,b) définissant une troisième catégorie de méthodes.

2.2.1 Notations

De manière générale, une image peut être représentée par une matrice de dimensions 3 où les deux premières représentent la position des pixels de l'image selon respectivement l'axe horizontal et l'axe vertical et la troisième le nombre de bandes spectrales.

En utilisant les conventions de la télédétection, on définit les matrices suivantes :

- la matrice $\mathbf{Y}$ de dimensions $\mathbb{R}^{M\times N\times T}$ correspond ici à l'image possédant l'information spatiale. Les deux premières dimensions M et N correspondent à la taille de l'image et donc à sa dimension spatiale. La dimension T est la dimension spectrale. Lorsque $T = 1$, $\mathbf{Y}$ est une image panchromatique.
- la matrice $\mathbf{X} = \{\mathbf{X}_k\}_{k=1,\dots,K}$ de dimensions $\mathbb{R}^{P\times Q\times K}$ représente l'image spectrale avec une résolution spatiale plus faible que l'image panchromatique. C'est-à-dire avec $P < M$ et $Q < N$. Cette matrice contient K bandes spectrales.
- La matrice $\tilde{\mathbf{X}}$ est une matrice intermédiaire de dimensions $\mathbb{R}^{M\times N\times K}$. Elle est appelée matrice suréchantillonnée, car elle contient les mêmes informations que la matrice $\mathbf{X}$ mais possède les mêmes dimensions spatiales que la matrice $\mathbf{Y}$.
- La matrice $\hat{\mathbf{X}}$ est la matrice obtenue après la fusion. Elle est de dimension $\mathbb{R}^{M\times N\times K}$. La résolution spatiale correspond alors à la résolution de l'image panchromatique et la résolution spectrale à celle de l'image spectrale.

3. On différencie ici les images multispectrales des images hyperspectrales par le nombre de bandes spectrales. Une image multispectrale possède quelques bandes spectrales tandis qu'une image hyperspectrale possède au moins une centaine de bandes spectrales.

2.2.2 Méthodes de substitution de composante (CS)

Cette classe de méthodes de pan-sharpening se base sur la réduction de la dimension spectrale des images spectrales. Ces méthodes se différencient cependant par la manière dont la réduction de dimension est effectuée. Une information globale issue de l'ensemble des bandes spectrales est récupérée. Parmi les méthodes appartenant à cette classe, on trouve : Intensity-Hue-Saturation (IHS) (Carper, 1990; Tu et al., 2001), Principal Component Analysis (PCA) (Chavez et al., 1991; Kwarteng and Chavez, 1989), Brovey Transform (BT) (Gillespie et al., 1987) et Gram-Schmidt (GS) (Laben and Brower, 2000; Aiazzi et al., 2007).

Une équation générique décrivant l'obtention d'une image avec de fortes résolutions spectrales et spatiales après fusion peut s'écrire de la manière suivante (Vivone et al., 2015) :

$$\hat{\mathbf{X}}_k = \tilde{\mathbf{X}}_k + g_k(\mathbf{P} - \mathbf{I}_L) \tag{2.1}$$

Avec $\mathbf{I}_L$ et g_k, les paramètres spécifiques aux différentes méthodes de la classe CS. La matrice $\mathbf{I}_L$ de dimension $\mathbb{R}^{M\times N\times T}$ est obtenue en réduisant la dimension spectrale de la matrice $\tilde{\mathbf{X}}_k$. Cette opération s'écrit :

$$\mathbf{I}_L = \sum_{i=1}^{K} \omega_i \tilde{\mathbf{X}}_k \tag{2.2}$$

Le vecteur $g_{i=1\ldots k}$ de dimension $\mathbb{R}^{1\times K}$ est communément appelé *fonction d'injection*. Cette fonction rétablit la dimension spectrale. Des poids peuvent être alors appliqués en fonction des bandes spectrales. Ici $\mathbf{P}$ est la matrice de dimensions $\mathbb{R}^{M\times N\times T}$ qui contient les valeurs ajustées de la matrice $\mathbf{Y}$. Ces valeurs sont obtenues par *appariement d'histogrammes* ou *histogram-matching*. Ceci permet de faire correspondre l'histogramme de l'image panchromatique avec l'histogramme de la matrice $\mathbf{I}_L$. L'appariement d'histogrammes peut se formuler de la manière suivante :

$$\mathbf{P} = (\mathbf{Y} - \mu_{\mathbf{Y}})\frac{\sigma_{\mathbf{I}_L}}{\sigma_{\mathbf{Y}}} + \mu_{\mathbf{I}_L} \tag{2.3}$$

Avec $\sigma_{\mathbf{I}_L}$ et $\sigma_{\mathbf{Y}}$ les écarts-types des matrices respectives $\mathbf{I}_L$ et $\mathbf{Y}$. $\mu_{\mathbf{I}_L}$ et $\mu_{\mathbf{Y}}$ correspondent aux moyennes respectivement des matrices $\mathbf{I}_L$ et $\mathbf{Y}$.

La table 2.1 précise les valeurs de chacun des paramètres ω_i et g_k spécifiques à chacune des méthodes CS. Les paramètres g_k et ω_i ont respectivement pour valeur 1 et 1/N pour la méthode IHS. Cela signifie que cette méthode réalise la moyenne de toutes les bandes spectrales. Les méthodes BT et GS réalisent la même opération que IHS pour former $\mathbf{I}_L$. Cependant, leur fonction d'injection g_k est différente. Pour la méthode BT, g_k

a pour valeur $\tilde{\mathbf{X}}_k/\mathbf{I}_L$. Pour la méthode GS, g_k prend la valeur de la covariance entre $\tilde{\mathbf{X}}_k$ et la matrice $\mathbf{I}_L$. La méthode PCA réalise une analyse en composantes principales. Seule la première composante principale $\mathbf{L}_{1,k}$ est récupérée. $\mathbf{L}_{1,k}$ est de dimension $\mathbb{R}^{1\times K}$. $\mathbf{I}_L$ est l'image formée par les scores $\mathbf{S}_1$ des pixels obtenus sur cette composante. Les loadings $\mathbf{L}_{1,k}$ de la première composante sont utilisés pour ensuite rétablir les dimensions de l'image hyperspectrale.

TABLE 2.1 – Spécifités des paramètres g_k et ω_i pour les méthodes de pan-sharpening appartenant à la classe CS.

Méthode	ω_i	g_k
IHS	$1/N$	1
PCA	$\mathbf{S}_1$	$\mathbf{L}_k$
BT	$1/N$	$\tilde{\mathbf{X}}_k/\mathbf{I}_L$
GS	$1/N$	$\mathrm{cov}(\tilde{\mathbf{X}}_k, \mathbf{I}_L)/\mathbf{I}_L$

2.2.3 Méthodes par Analyse Multi-Résolution (MRA)

Les méthodes appartenant à la classe d'analyse multi-résolution (MRA) appliquent un filtre à l'image panchromatique. Parmi les méthodes appartenant à cette classe, on trouve entre autres : High-Pass-Filter (HPF) (Carper, 1990), Smoothing Filter based-Intensity Modulation (SFIM) (Liu, 2000), Generalized Laplacian Pyramid with Modulation Transfer Function/with High-Pass-Modulation (MTF-GLP, MTF-GLP-HPM) (Alparone et al., 2007; Aiazzi et al., 2003) et Guided Filtering in PCA domain (GFPCA) (Liao et al., 2015).

Comme précédemment (Eq. 2.3), la formulation de cette catégorie de méthodes peut s'écrire selon la littérature (Vivone et al., 2015) de la manière suivante :

$$\hat{\mathbf{X}}_k = \tilde{\mathbf{X}}_k + g_k(\mathbf{P} - \mathbf{P}_L) \tag{2.4}$$

Avec $\mathbf{P}_L$, le résultat issu de l'application d'un filtre sur l'image panchromatique

Les méthodes se différencient par les filtres utilisés et le terme d'injection g_k. Ces spécificités sont données dans la table 2.2.

Pour les méthodes HPF et SFIM, le filtre utilisé est le filtre linéaire 'Box Filter'. Les termes g_k utilisés pour HPF et SFIM sont respectivement 1 et $\tilde{\mathbf{X}}_k/\mathbf{P}_L$. Les méthodes MTF-GLP et MTF-GLP-HPM utilisent un filtre appelé Modulation Transfer Function (MTF). Ce filtre modélise la fonction de réponse en fréquence spatiale du capteur spectral. La méthode GFPCA est similaire à la méthode PCA de la classe CS, mais utilise un filtre guidé. Ce filtre utilise l'information spatiale contenue dans l'image panchromatique pour préserver les contours.

TABLE 2.2 – Spécifités des paramètres des méthodes de pan-sharpening appartenant à la catégorie MRA.

Méthode	Filtre	g_k
HPF	Box Filter	1
SFIM	Box Filter	$\tilde{\mathbf{X}}_k/\mathbf{P}_L$
MTF-GLP	MTF	1
MTF-GLP-HPM	MTF	$\tilde{\mathbf{X}}_k/\mathbf{P}_L$
GFPCA	Guided Filter	$\mathbf{L}_k$

2.2.4 Méthodes basées sur le démélange spectral

D'autres méthodes ont été créées pour de la fusion de données hyperspectrales avec des données multispectrales ou panchromatiques. Ces méthodes de pan-sharpening peuvent être également appelées méthodes de multi-sharpening (Benhalouche et al., 2015). Ces méthodes reposent sur des techniques de démélange spectral. Le **mélange** spectral est défini comme un spectre qui peut se décomposer en plusieurs spectres issus de différents éléments appelés aussi matériaux purs. Un système de spectres mélangés x_i peut alors être composé de p éléments individuels (Ruckebusch, 2016). Le mélange peut se modéliser avec l'équation suivante :

$$x_i = \sum_{j=1}^{p} w_j h_{i,j} + \epsilon_i \tag{2.5}$$

Où w_j contient les spectres de matériaux purs et est appelé *endmembers*. $h_{i,j}$ est un terme appelé *abondances* et correspond aux proportions de chaque composant de w_j. ϵ_i est l'erreur résiduelle du modèle.

Le démélange ou *unmixing* correspond au problème inverse consistant à séparer l'information spectrale x_i contenue dans un pixel en une collection d'éléments individuels. La résolution d'un tel problème inverse passe généralement par une étape de modélisation de ce phénomène.

Parmi les modèles existants, le modèle de mélange linéaire ou *Linear mixing model* (LMM) reste simple, car le spectre est représenté par une combinaison linéaire des différents composants w_j. Des modèles de mélange non linéaire ou *Non linear mixing model* (NLMM) existent et sont plus représentatifs des milieux diffusants. Parmi les méthodes non-linéaires, les méthodes de la famille des modèles de mélanges bilinéaires proposent d'ajouter un terme d'interaction de second ordre entre les spectres de matériaux purs (Dobigeon et al., 2016) : $\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \beta_{i,j} w_i \odot w_j$ avec $\odot$ le produit matriciel de Hadamard (terme par terme) et $\beta_{i,j}$ le paramètre d'interaction entre les composants w_i et w_j.

Nous décidons d'utiliser deux méthodes linéaires : CNMF et HySure qui apparaissent intéressantes et performantes (Loncan et al., 2015)

CNMF

La factorisation couplée de matrices non-négatives ou Couple Nonnegative Matrix Factorization (CNMF) est une méthode basée sur la technique de démélange. Cette méthode utilise la factorisation de matrices non-négatives ou nonnegative matrix factorization (NMF) (Liu et al., 2010). L'objectif de cette méthode est de reconstruire une image $\mathbf{Z}$ ayant la résolution spectrale de l'image hyperspectrale $\mathbf{X}$ et la résolution spatiale de la matrice $\mathbf{Y}$.

Dans cette partie, les matrices utilisées sont des matrices de deux dimensions où la première dimension correspond à la dimension spectrale et la seconde à la dimension spatiale. En reprenant les mêmes notations que précédemment 2.2.1, c'est-à-dire : M, N, les dimensions spatiales élevées ; P, Q les dimensions spatiales faibles ; K la dimension spectrale élevée et T la dimension spectrale faible. La matrice $\mathbf{Z}$ est donc de dimension $\mathbb{R}^{K\times L_{M,N}}$ où $L_{M,N}$ est la dimension correspondant à la taille du vecteur contenant les pixels de l'image. La matrice $\mathbf{X}$ est de dimensions $\mathbb{R}^{K\times L_{P,Q}}$ et la matrice $\mathbf{Y}$ est de dimension $\mathbb{R}^{T\times L_{M,N}}$ Les images $\mathbf{X}$ et $\mathbf{Y}$ sont modélisées à partir de la matrice $\mathbf{Z}$. Cette modélisation fait intervenir deux matrices :

- la matrice $\mathbf{S}$ de dimension $\mathbb{R}^{L_{M,N}\times L_{P,Q}}$ traduit la dégradation spatiale de la caméra hyperspectrale. Cette matrice est aussi appelée Point Spread Function (PSF).
- la matrice $\mathbf{R}$ de dimension $\mathbb{R}^{T\times K}$ représente la dégradation spectrale de la caméra multispectrale.

À partir de ces deux matrices, les images $\mathbf{X}$ et $\mathbf{Y}$ peuvent être modélisées de la manière suivante :

$$\mathbf{X} = \mathbf{ZS} + \mathbf{E}_s \tag{2.6}$$

$$\mathbf{Y} = \mathbf{RZ} + \mathbf{E}_r \tag{2.7}$$

Avec $\mathbf{E}_s$ et $\mathbf{E}_r$ les erreurs résiduelles de la modélisation de la dégradation respectivement spatiale et spectrale. Ici la modélisation fait intervenir l'image souhaitée, c'est-à-dire la matrice $\mathbf{Z}$. Cette même matrice est ensuite modélisée par le modèle de mélange linéaire suivant l'équation 2.5. Cette équation peut se réécrire de manière matricielle avec l'équation suivante :

$$\mathbf{Z} = \mathbf{WH} + \mathbf{N} \tag{2.8}$$

Avec $\mathbf{N}$ la matrice contenant l'erreur résiduelle, $\mathbf{W}$ la matrice de signature spectrale de dimensions $\mathbb{R}^{K\times O}$ et $\mathbf{H}$, la matrice d'abondance de dimension $\mathbb{R}^{O\times L_{M,N}}$ où O correspond au nombre d'éléments individuels. En utilisant l'équation 2.8, il est possible de réécrire les équations 2.6 et 2.7 de la manière suivante :

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}_h \tag{2.9}$$

$$\mathbf{Y} \approx \mathbf{W}_m\mathbf{H} \tag{2.10}$$

Où $\mathbf{H}_h$ est la matrice d'abondance dégradée spatialement avec $\mathbf{H}_h = \mathbf{H}\mathbf{S}$. $\mathbf{W}_m$ est la matrice de signature spectrale dégradée spectralement obtenue selon l'équation : $\mathbf{W}_m = \mathbf{R}\mathbf{W}$. Les matrices $\mathbf{H}_h$ et $\mathbf{W}_m$ sont de dimensions $\mathbb{R}^{O\times L_{P,Q}}$ et $\mathbb{R}^{T\times O}$ respectivement.

Les équations 2.8, 2.9 et 2.10 constituent un système d'équations. La résolution de la méthode CNMF passe par de l'optimisation minimisant alors $\|\mathbf{X} - \mathbf{W}\mathbf{H}_h\|^2$ et $\|\mathbf{Y} - \mathbf{W}_m\mathbf{H}\|^2$. La norme utilisée $\|.\|^2$ est la norme de Frobenius. Les deux équations 2.9 et 2.10 sont obtenues en utilisant la NMF.

HySure

La méthode HySure (Simões et al., 2015; Simoes et al., 2014) pose également un problème d'optimisation en utilisant une fonction objectif convexe [4]. Cette méthode contient plusieurs étapes, dont une étape de réduction de dimension spectrale de l'image hyperspectrale et une étape de résolution d'un problème de régularisation. L'étape de réduction de dimensions peut s'écrire de la manière suivante :

$$\mathbf{Z} = \mathbf{E}\mathbf{X} \tag{2.11}$$

Ici, la matrice $\mathbf{E}$ de dimensions $\mathbb{R}^{U\times K}$ avec $U < K$ peut être obtenue selon des principes d'extraction non supervisée des matériaux purs comme la méthode *Vertex Component Analysis* (VCA) (Nascimento and Dias, 2005).

Le terme de régularisation est défini comme un vecteur de variation totale $\phi(\mathbf{X})$ (Bresson and Chan, 2008). Ce terme permet de favoriser les solutions dites *lisses par morceaux* et favorise la continuité spectrale entre les pixels de l'image. Le terme de régularisation s'écrit selon l'équation suivante :

$$\phi(\mathbf{X}) = \sum_{j=1}^{N}\sqrt{\sum_{i=1}^{M}\{[(\mathbf{X}\mathbf{D}_h)_{i,j}]^2 + [(\mathbf{X}\mathbf{D}_v)_{i,j}]^2\}} \tag{2.12}$$

4. Les fonctions convexes ont des propriétés très importantes en optimisation. Elles permettent de trouver systématiquement un minimum en évitant les problèmes liés aux minima locaux/globaux.

Où les matrices $\mathbf{D}_h$ et $\mathbf{D}_v$ sont les matrices qui contiennent les différences discrètes horizontales et verticales. En utilisant les mêmes notations employées pour la méthode CNMF, la fonction à minimiser peut s'écrire de manière suivante :

$$f = \frac{1}{2}\|\mathbf{X} - \mathbf{ZS}\|^2 + \frac{\lambda_m}{2}\|\mathbf{Y} - \mathbf{RZ}\|^2 + \lambda_\phi \phi(\mathbf{D}_h, \mathbf{D}_v) \tag{2.13}$$

Où λ_ϕ et λ_m représentent les poids des différents termes à ajuster.

2.3 Matériel et méthodes

2.3.1 Démarche générale

Les algorithmes présentés précédemment sont testés selon le protocole de Wald (Wald et al., 1997). Ce protocole peut se décomposer en trois étapes. La première est de partir d'une image de référence avec une résolution spatiale/spectrale souhaitée et de la dégrader dans un premier temps spatialement. Dans un second temps, cette image de référence est dégradée spectralement. Ce procédé permet d'obtenir deux images. Une avec une forte résolution spectrale/faible résolution spatiale et une seconde avec une faible résolution spectrale/forte résolution spatiale. La deuxième étape est de reconstruire une image avec les méthodes de pan-sharpening en utilisant les images dégradées obtenues lors de l'étape précédente. La troisième étape est de caractériser l'erreur de reconstruction entre l'image de référence et l'image obtenue par pan-sharpening. On utilise pour ce faire des indices d'erreur dédiés et décris dans la section suivante.

2.3.2 Indices utilisés pour l'évaluation de l'erreur de reconstruction

Nous décidons d'utiliser trois indices pour caractériser l'erreur de la reconstruction. Les deux premiers sont complémentaires et permettent de dissocier respectivement les erreurs de nature spatiale et celles de nature spectrale. Le troisième indice est un indice global qui intègre les deux sources d'erreurs. Ces trois indices sont généralement utilisés pour comparer les performances des méthodes de pan-sharpening (Vivone et al., 2015; Loncan et al., 2015). Les indices utilisés sont les suivants :

- Corrélation croisée ou Cross Correlation (CC) : cet indice calcule la corrélation entre deux images. Il est très utilisé en imagerie notamment pour rechercher des motifs spatiaux. Dans notre cas, nous l'utilisons pour comparer deux images hyperspectrales. Cet indice calcule la moyenne des valeurs de corrélation obtenues

pour chacune des bandes spectrales.

$$\mathrm{CC}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{M \times N} (\mathbf{X}_i - \mu_{\mathbf{X}})(\hat{\mathbf{X}}_i - \mu_{\hat{\mathbf{X}}})}{\sqrt{\sum_{i=1}^{M \times N} (\mathbf{X}_i - \mu_{\mathbf{X}})^2 + \sum_{i=1}^{M \times N} (\hat{\mathbf{X}}_i - \mu_{\hat{\mathbf{X}}})^2}} \right) \tag{2.14}$$

Où $\mu_{\mathbf{X}}$ et $\mu_{\hat{\mathbf{X}}}$ sont les moyennes respectives des matrices $\mathbf{X}$ et $\hat{\mathbf{X}}$. Cet indice varie de 0 à 1. Lorsqu'il est proche de 0, cela indique qu'il y a une déformation géométrique importante entre les deux images. Lorsque la valeur est proche de 1, il y a très peu de déformation.

- Spectral Angle Mapper (SAM) (Kruse et al., 1993; Yuhas et al., 1992) : Le SAM est l'angle formé entre deux spectres $\mathbf{s}_1$ et $\mathbf{s}_2$ dans l'espace vectoriel défini par leurs longueurs d'onde. Il est exprimé en degrés ou en radians et se calcule de la manière suivante :

$$\mathrm{SAM}(\mathbf{s}_1, \mathbf{s}2) = \arccos \left(\frac{\langle \mathbf{s}_1, \mathbf{s}_2 \rangle}{\|\mathbf{s}_1\| \, \|\mathbf{s}_2\|} \right) \tag{2.15}$$

  Avec $\|.\|$ la norme ℓ^2. Cet indice décrit la similarité spectrale entre deux spectres. Lorsque l'angle est proche de 0°, les spectres sont similaires. Lorsque les angles sont grands, les spectres ne se ressemblent pas. La valeur de cet indice est indépendante de l'intensité du signal. Dans notre étude, le SAM est tout d'abord calculé pour chaque pixel des images. Un indice final est obtenu en moyennant l'ensemble des angles obtenus pour chaque pixel des images.

- L'erreur quadratique moyenne ou Root Mean Square Error (RMSE) est une différence entre les valeurs prédites par un modèle et les valeurs observées. Cet indice est calculé de la manière suivante :

$$\mathrm{RMSE}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{\left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F}{\sqrt{MNK}} \tag{2.16}$$

  Où $\|.\|_F$ est la norme de Frobenius. Cet indice intègre à fois les erreurs de nature spatiale et spectrale. Un RMSE proche de zéro indique une erreur très faible entre les deux images. Plus la valeur est grande, plus l'écart entre les deux images est élevé.

2.3.3 Image hyperspectrale de référence

Une image hyperspectrale a été utilisée comme image de référence. Cette image a été acquise sur une scène de betterave sucrière (Fig. 2.1) à l'aide d'une caméra hyperspectrale HySpex VNIR 1600 (NEO, Norvège) disposant de 160 bandes spectrales dans la gamme

FIGURE 2.1 – Représentation couleur de l'image hyperspectrale de référence : betterave sucrière.

du visible et du proche infrarouge (409- 987 nm) positionnée à une hauteur d'environ 1 mètre au-dessus de la canopée. L'image obtenue présente une résolution spatiale élevée de 1,15 mm/pixel.

2.3.4 Cas d'étude

Les méthodes de pan-sharpening présentées précédemment, à savoir : IHS, PCA, BT, GS, HPF, SFIM, MTF-GLP, MTF-GLP-HPM, CNMF (avec un nombre de matériaux purs fixé à 9) et HySure seront comparées dans deux cas d'étude : le premier est l'étude de l'influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction. Le deuxième est l'étude de l'influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction. Pour ces deux cas d'étude, les résultats seront évalués en utilisant les indices présentés en section 2.3.2. L'objectif est alors de sélectionner la méthode qui fournit les meilleurs résultats sur la base de ces indices.

Influence de la résolution spatiale de l'image spectrale

Dans cette partie, nous cherchons à étudier l'influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction des images obtenues après l'utilisation des méthodes de pan-sharpening décrites en section 2.2. Pour cela, nous adaptons le protocole de Wald en faisant varier la résolution spatiale de l'image spectrale.

À partir de l'image hyperspectrale de référence, une image panchromatique de dimension 1 et de résolution spatiale élevée est extraite. L'image panchromatique est obtenue

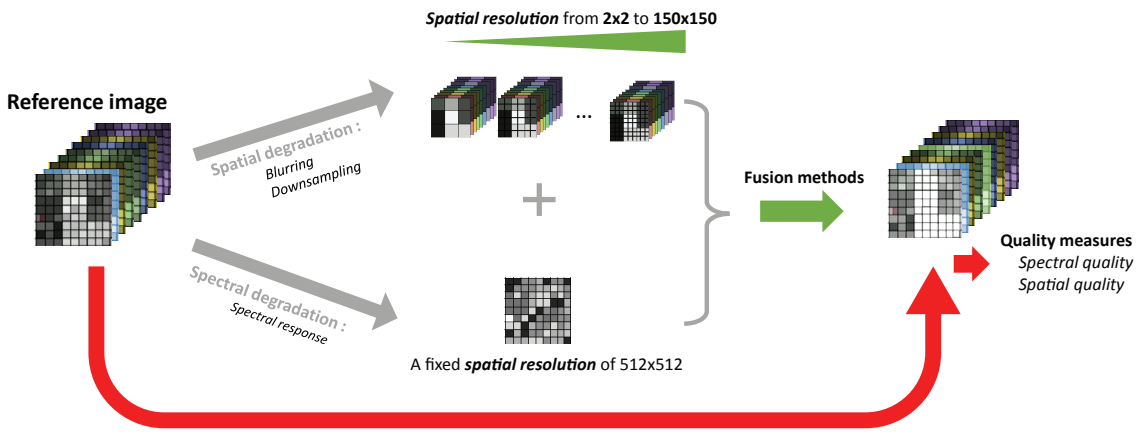


FIGURE 2.2 – Méthodologie pour évaluer l'influence de la résolution spatiale de l'image spectrale sur les qualités de reconstruction des méthodes de pan-sharpening.

en utilisant une seule bande spectrale définie à $735\ \mathrm{nm}$ permettant ici de bien différencier la végétation du sol. D'autre part, l'information spectrale a été extraite par dégradation spatiale dans une grille régulière de NxN pixels où N appartient à l'intervalle [2, 150]. Le spectre assigné au nouveau pixel est la moyenne des spectres contenus dans l'image hyperspectrale de référence. Une image spectrale contient un nombre de spectres défini par N^2. Les spectres sont obtenus en utilisant l'ensemble de la plage spectrale disponible de l'image de référence. L'ensemble de la procédure est illustrée sur la figure 2.2.

Influence de la résolution spatiale de l'image panchromatique

Dans cette partie, on étudie l'influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction. Comme précédemment (Section 2.3.4), la procédure se base sur le protocole de Wald. Ici la résolution de l'image spectrale est fixée à 32x32 tandis que la résolution spatiale (défini par MxM) de l'image panchromatique varie de 32x32 à 512x512. Les mêmes indices CC, SAM et RMSE sont calculés pour évaluer la qualité de la reconstruction. La procédure est illustrée avec la figure 2.3.

2.4 Résultats et Discussions

2.4.1 Influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction et sur le temps de calcul

Nous appliquons ainsi l'approche décrite en section 2.3.4 à savoir : (1) la dégradation de la résolution spectrale de l'image hyperspectrale de référence pour obtenir une image

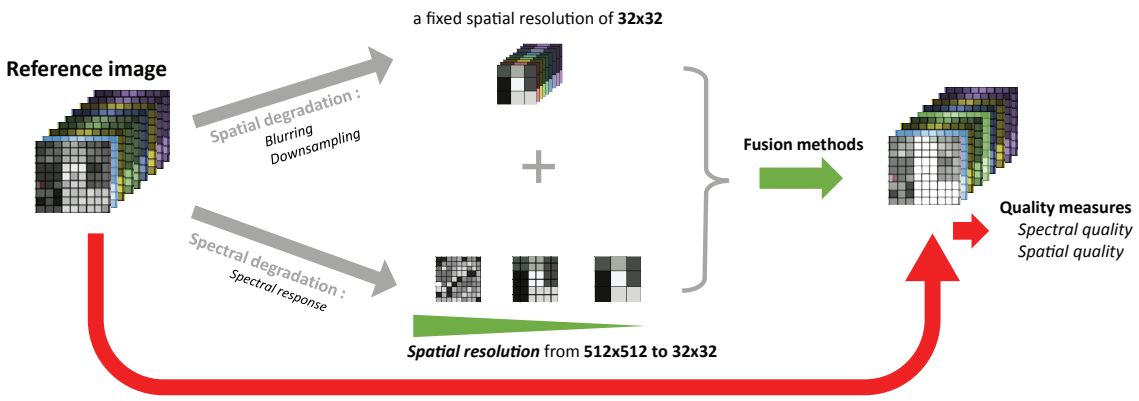


FIGURE 2.3 – Méthodologie pour évaluer l'influence de la résolution spatiale de l'image panchromatique sur les qualités de reconstruction des méthodes de pan-sharpening.

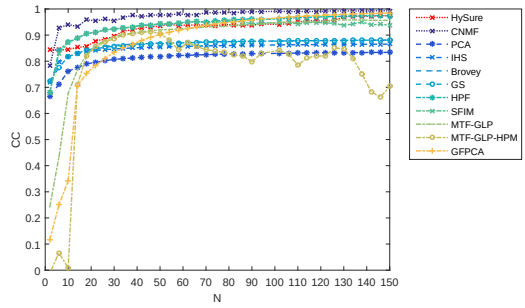
panchromatique de résolution 512x512 (2) la variation de la résolution spatiale de l'image spectrale, (3) l'application des approches de pan-sharpening et comparaison de l'image reconstruite avec l'image de référence en utilisant les indices CC, SAM et RMSE.

Qualité de reconstruction

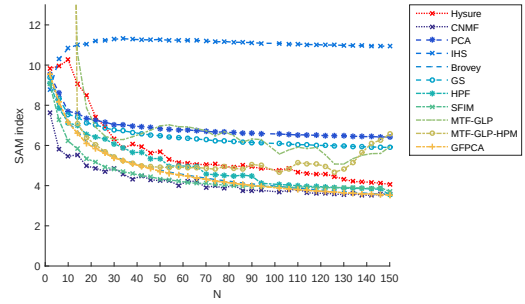
Indices moyennés Les figures 2.4 montrent l'évolution des valeurs des indices obtenus pour l'ensemble des méthodes testées en fonction de la résolution spatiale (NxN) de l'image spectrale.

La courbe 2.4a montre l'évolution de l'indice CC. Pour l'ensemble des méthodes sauf l'approche MTF-GPL, cet indice augmente rapidement pour des valeurs de N comprises entre 2 et 20 puis se stabilise pour des valeurs de N supérieurs ou égales à 30. Les valeurs de CC obtenues pour l'approche MTF-GLP fluctuent par contre énormément en particulier pour des valeurs de N supérieures à 100. Les dynamiques des courbes de l'évolution de cet indice en fonction de N et pour l'ensemble des méthodes testées sont différentes : la méthode CNMF obtient des valeurs supérieures aux autres méthodes, et ce pour tout N. Les valeurs sont supérieures à 0.9 à partir de N=6. Les courbes CC pour les méthodes SFIM et HPF se superposent pour des valeurs de N comprises entre 2 et 80. La méthode GFPCA présente des valeurs très faibles pour des valeurs de N inférieure à 60. Pour la méthode HySure, les valeurs de CC sont globalement élevées quel que soit N et augmentent lentement et de façon constante selon N pour atteindre des valeurs proches de la méthode CNMF et GFPCA.

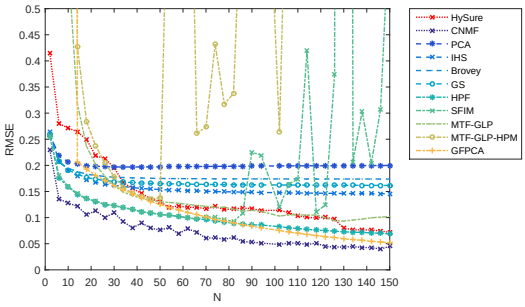
CC caractérise la qualité de reconstruction spatiale. Sur la base de cet indice, les méthodes CS apparaissent les moins performantes et ceci quel que soit la valeur de N.



(a)



(b)



(c)

FIGURE 2.4 – Evolution des indices pour les différentes méthodes de pan-sharpening testées en fonction de la résolution spatiale N : (a) CC, (b) SAM, (c), RMSE.

Pour une résolution forte de l'image spectrale (N>100), les méthodes CNMF, HySure, GFPCA et HPF sont les plus performantes. Pour des résolutions spatiales faibles (N<50), la méthode CNMF semble la plus appropriée.

La courbe 2.4b montre l'évolution de l'indice SAM en fonction de N pour l'ensemble des méthodes de pan-sharpening testées. L'évolution de la valeur de cet indice pour l'approche IHS se différencie nettement de celles des autres approches testées. Ainsi, pour l'ensemble des approches testées hormis l'approche IHS, plus N augmente plus l'indice SAM est faible indiquant une similarité spectrale accrue entre l'image de référence et l'image reconstruite. Pour des valeurs de N supérieures à 40, les valeurs de SAM les plus faibles sont obtenues pour CNMF et SFIM. Les valeurs les plus élevées sont obtenues

pour les méthodes HySure et IHS. Pour N>100, les valeurs de SAM les plus faibles sont obtenues pour les méthodes HySure, CNMF, GFPCA et HPF et SFIM. L'approche CNMF présente des valeurs de SAM inférieures aux autres méthodes, quelle que soit la résolution spatiale N.

La figure 2.4c montre l'évolution des valeurs des erreurs RMSE pour chacune des méthodes en fonction de la résolution spatiale de l'image spectrale. Lorsque N augmente, les valeurs de RMSE des méthodes MTF-GLP-HPM et SFIM fluctuent énormément tandis qu'elles se stabilisent rapidement entre 0.15 et 0.20 à partir de N = 10 pour les méthodes PCA,GS, BT et IHS. Pour les autres méthodes, les valeurs de RMSE diminuent progressivement : À faible résolution (N<40), les méthodes HySure, MTF-GLP, et GFPCA présentent des valeurs RMSE plus faibles que les autres approches. La méthode CNMF présente des valeurs de RMSE inférieures aux autres approches quel que soit la valeur de N.

Avec des résolutions spatiales élevées (N>100), les méthodes CNMF et GFPCA présentent les meilleures performances. Pour des résolutions spatiales faibles (N<50), ce sont les méthodes CNMF, SFIM et HPF.

Détails des indices SAM et CC pour les meilleures méthodes de chaque classe L'étude précédente s'est intéressée à l'évaluation de la qualité globale de reconstruction de l'image. En effet, les indices SAM et CC ont été moyennés respectivement par pixels et par longueur d'onde. Une étude plus approfondie consiste alors à étudier les valeurs de SAM pixel à pixel et les valeurs de CC en fonction des longueurs d'onde. Pour cela, nous fixons alors la résolution spatiale de l'image spectrale à 32x32.

Valeurs de l'indice SAM pixel à pixel Les images, en figure 2.5, montrent alors les valeurs de SAM par pixel pour chacune des trois meilleures méthodes de chaque classe. Pour l'ensemble des méthodes, des valeurs faibles de SAM sont obtenues pour des pixels correspondant aux feuilles de la plante. En revanche, pour la méthode BT (Fig. 2.5a), les valeurs de SAM sont élevées pour des pixels situés à l'interface feuille/sol. Ce qui est assez surprenant c'est que les images correspondant à l'indice SAM des méthodes BT (Fig. 2.5a) et SFIM (Fig. 2.5b) semblent pixellisées. En revanche, l'image reconstruite par CNMF (Fig. 2.5c) ne présente pas de pixellisation ou agrégation entre les pixels.

La pixellisation de l'image SAM traduit ici une erreur de reconstruction de nature spatiale. En effet cette discontinuité correspond ici à la résolution de l'image spectrale fixée à 32x32. Ce phénomène montre effectivement que les méthodes de pan-sharpening appartenant aux classes CS et MRA ne sont pas adaptées pour des résolutions spatiales aussi éloignées entre les deux images. En effet, ces méthodes sont généralement utilisées

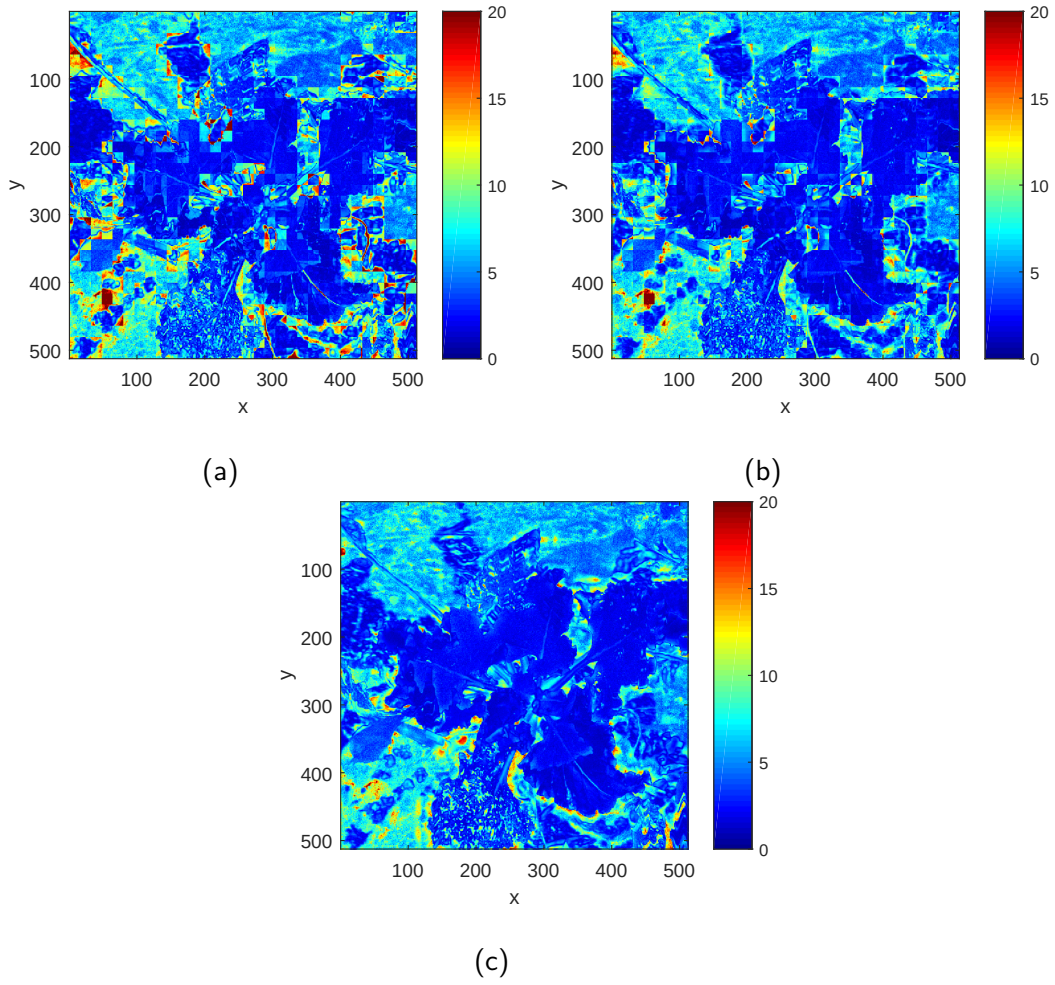


FIGURE 2.5 – Qualité de reconstruction pixel à pixel selon l'indice SAM pour différentes méthodes de pan-sharpening : (a) BT, (b) SFIM, (c), CNMF.

pour des images satellites. Dans ce cas, le ratio entre résolutions spatiales de l'image multispectrale et panchromatique varie aux alentours de 1/4 ou 1/2. Dans notre cas d'étude, ce ratio varie entre 1/4 (N=128) à 1/256 (N=2).

Valeurs de l'indice CC par longueur d'onde La figure 2.6 montre l'indice CC obtenu longueur d'onde par longueur d'onde. Entre 400 nm et 700 nm, les valeurs de CC obtenues sont élevées pour les 3 méthodes choisies. En revanche, ces méthodes se distinguent dans le proche infra-rouge. Pour la méthode CNMF, l'indice CC est supérieur aux deux autres approches pour l'ensemble des longueurs d'onde et est relativement stable avec une variation inférieure à 5%. Au contraire, l'indice CC pour la méthode BT varie de 0.94 à 0.71 selon la longueur d'onde.

Contrairement aux deux méthodes SFIM et CNMF, la qualité de reconstruction spatiale de la méthode BT dépend énormément de la longueur d'onde. Cette méthode permet difficilement de reconstruire des images avec une qualité spatiale suffisante dans

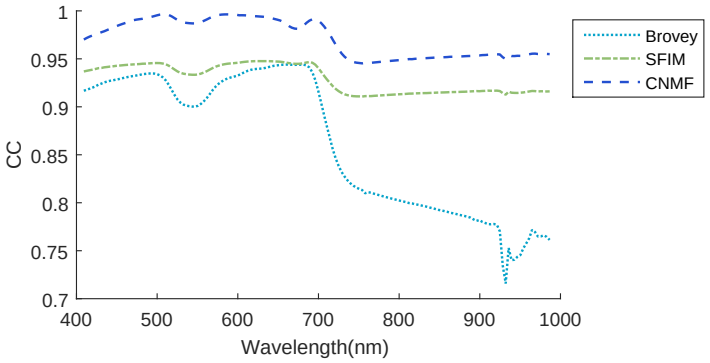
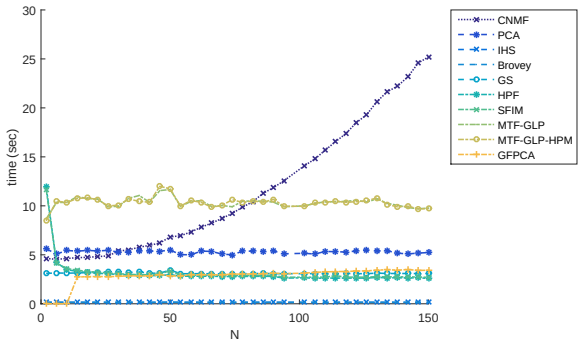


FIGURE 2.6 – Le coefficient CC par longueur d'onde.

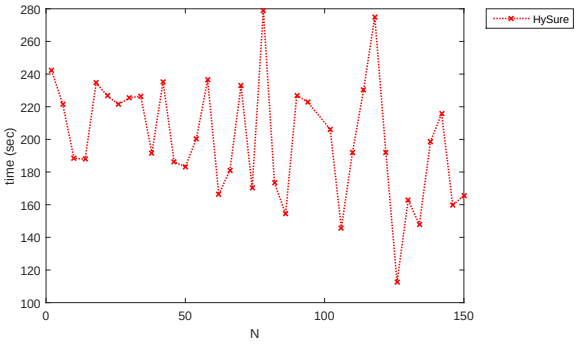
le proche-infrarouge.

Temps de calcul des différentes méthodes de pan-sharpening testées en fonction de N Un autre élément à prendre en compte pour la sélection d'une approche de fusion est le temps de calcul. Pour chaque méthode, les temps de calcul en fonction de la résolution spatiale de l'image spectrale ont été calculés (Fig. 2.7a et 2.7b). Les temps de calcul de la méthode HySure (Fig. 2.7b) sont très supérieurs à ceux obtenus pour les autres méthodes avec des durées qui fluctuent entre 120 secondes et 280 secondes. Seules les approches CNMF et HySure ont un temps de calcul qui évolue en fonction de N. Les temps de calcul les plus faibles sont obtenus avec la méthode IHS avec en moyenne une durée de 0.15 secondes et cela quel que soit N. La reconstruction de l'image via les méthodes SFIM, HPF, GFPCA et GS durent environ 3 secondes. La méthode PCA a un temps de calcul de 5 secondes. Les méthodes MTF-GLP et MTF-GLP-HPM reconstruisent l'image en 10 secondes.

La résolution spatiale de l'image spectrale n'impacte pas les temps de calcul des méthodes CS et MRA. En revanche, elle impacte les temps de calcul des méthodes CNMF et HySure. Dans ce cas, plus la résolution spatiale est élevée, plus le temps de calcul est élevé. En prenant en compte les temps de calcul et la qualité de reconstruction via les indices CC, SAM et RMSE, la méthode CNMF apparait comme la plus performante.



(a)



(b)

FIGURE 2.7 – Comparaison des durées pour reconstruire une image en fonction de N pour : (a) l'ensemble des méthodes sauf HySure (b) la méthode HySure.

2.4.2 Influence de la résolution spatiale de l'image panchromatique sur la qualité de reconstruction et temps de calcul

Indices moyennés Les figures 2.8a, 2.8b et 2.8c présentent les courbes d'évolution des indices CC, SAM et RMSE selon la résolution spatiale de l'image panchromatique (MxM) pour chacune des méthodes de pan-sharpening testées. Pour des valeurs M>100, les valeurs de ces trois indices varient très peu pour l'ensemble des méthodes. En revanche lorsque M<100, les valeurs diminuent pour les indices CC et augmentent pour le RMSE. Pour l'indice SAM, les valeurs ne varient pas pour la majorité des méthodes. Seules les valeurs de la méthode HySure augmentent lorsque M<300.

Hormis la méthode HySure, la résolution spatiale de l'image panchromatique a très peu d'influence lorsque $M > 100$. En revanche, lorsque la résolution spatiale de l'image panchromatique est très proche de la résolution spatiale de l'image spectrale (N=32), la qualité de reconstruction est moins bonne.

Temps de calcul des différentes méthodes de pan-sharpening testées en fonction de M La figure 2.9 montre les temps de calcul pour chacune des méthodes de

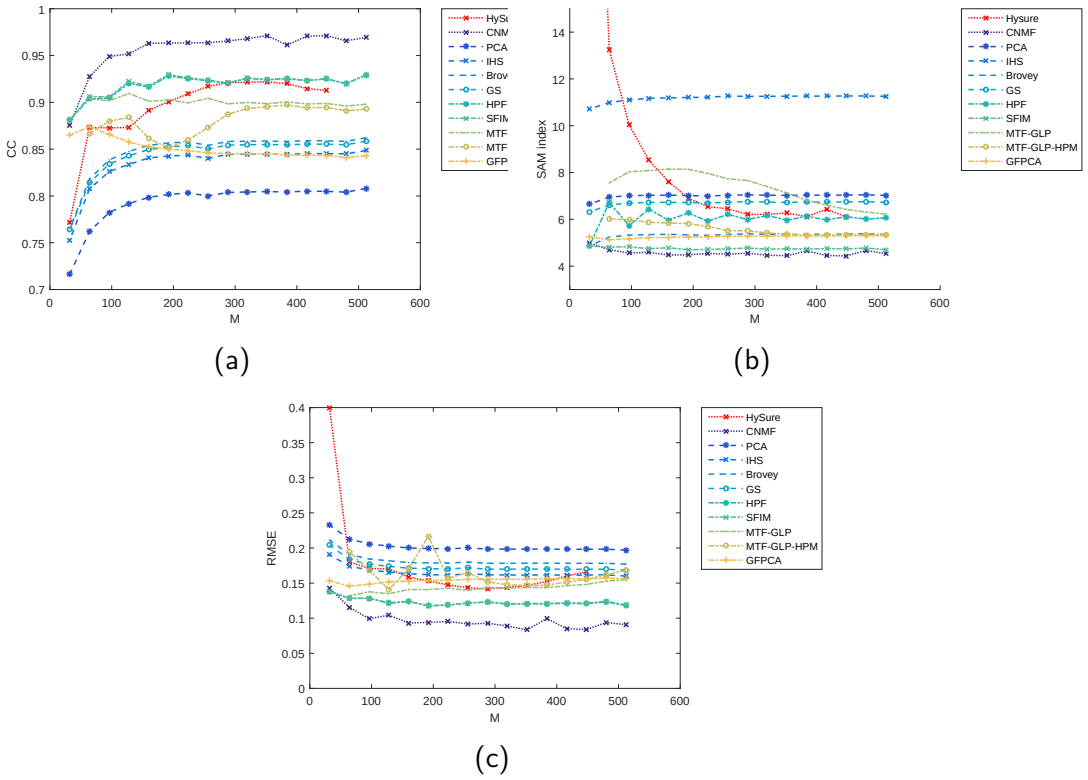


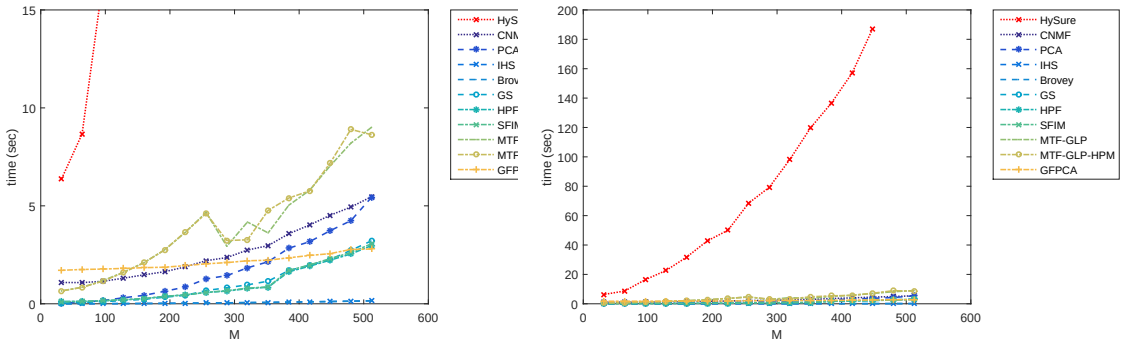
FIGURE 2.8 – Comparaison des méthodes de pan-sharpening en fonction de M suivant les indices : (a) CC, (b) RMSE, (c), SAM (deg).

pan-sharpening testée. Les temps de calcul de la méthode HySure (Fig. 2.9b) sont très supérieurs au temps de calcul des autres méthodes et augmentent lorsque M augmente.

Pour les autres méthodes (Fig. 2.9a), plus la résolution spatiale de l'image panchromatique est élevée, plus les temps de calcul augmentent pour l'ensemble des méthodes de pan-sharpening. Les méthodes qui présentent les temps de calcul les plus faibles (IHS, BT) sont celles qui sont les moins complexes et qui ne requièrent que des calculs simples. De façon inverse, l'approche HySure présente des temps de calcul beaucoup plus élevés. Cette approche est plus complexe à mettre en oeuvre et fait intervenir des boucles d'optimisation. Une augmentation de la résolution spatiale de l'image panchromatique entraîne directement une augmentation considérable du temps de calcul.

2.4.3 Conclusion partielle

Cette étude a permis de comparer un ensemble de méthodes de pan-sharpening et de tester l'influence de la résolution spatiale de l'image spectrale sur la qualité de reconstruction. Une méthode basée sur le démélange spectral (CNMF) s'est montrée particulièrement performante. L'image reconstruite via cette méthode est la plus fidèle à l'image de référence d'un point de vue spectral et spatial selon les indices choisis.



(a) (b)

FIGURE 2.9 – Comparaison de la durée de calculs des méthodes de pan-sharpening en fonction de la résolution spatiale M de l'image panchromatique avec : (a) une échelle adaptée à la majorité des méthodes, (b) selon une échelle adaptée à HySure.

La résolution spatiale de l'image panchromatique a très peu d'impact sur les images reconstruites quel que soit le ratio résolution spatiale image panchromatique/image spectrale. Le seul impact est uniquement d'ordre opérationnel. En effet, les temps de calcul des deux méthodes basées sur le démélange (HySure et CNMF) peuvent être réduits en diminuant la résolution spatiale de l'image panchromatique. Un compromis doit alors être fixé entre le temps de calcul et la résolution spatiale finale. Pour la suite de l'étude, la méthode utilisée sera CNMF avec une résolution spatiale de l'image spectrale de 32x32.

2.5 Problème de recouvrement

Généralement les deux images utilisées par les méthodes de pan-sharpening se superposent. Ces images n'ont donc pas de problème d'appariement de pixels. Un pixel de l'image spectrale correspond alors à plusieurs pixels de l'image panchromatique. Nous appelons ces pixels : les pixels appariés.

En revanche, deux capteurs combinés peuvent ne pas mesurer systématiquement les mêmes surfaces. Autrement dit, des pixels de l'image panchromatique peuvent ne pas faire partie de l'image spectrale. Nous appelons ces pixels : les pixels non-appariés. Nous proposons dans cette partie deux méthodes permettant d'attribuer de l'information spectrale pour des pixels non-appariés à partir de pixels appariés.

2.5.1 Méthode proposée pour le manque de recouvrement

Nous proposons ici d'adapter la méthode CNMF. La première étape est donc de récupérer les pixels appariés de l'image panchromatique et de l'image spectrale. Deux images nommées pseudo-images sont formées à partir de ces pixels appariés. La pseudo-image

panchromatique est construite à partir des pixels appariés de l'image panchromatique. La deuxième pseudo-image est formée à partir des pixels appariés de l'image spectrale. La méthode CNMF est ensuite appliquée sur les deux pseudo-images pour construire une image hyperspectrale.

L'image hyperspectrale construite et la pseudo-image panchromatique servent de jeu d'entraînement (appelé aussi *training set*). À l'aide de ce jeu d'entraînement, nous proposons donc deux méthodes permettant d'attribuer des spectres aux pixels non-appariés.

Approche CNMF-KNN

Ici, une méthode d'apprentissage supervisée est utilisée, la méthode des k-plus-proches voisins (*k-nearest neighbors* ou KNN) (Sharaf et al., 1986). C'est une méthode de classification qui s'appuie sur la notion de distances entre observations. Un de ses avantages est qu'elle est adaptée pour des problèmes non-linéaires (Bakeev, 2010). Pour cette classification, il est nécessaire d'avoir un jeu d'entraînement. Nous utiliserons ici celui constitué des couples de pixels issus de la pseudo-image panchromatique et de l'image reconstruite par la méthode CNMF. Pour attribuer un spectre à un pixel, la première étape est d'effectuer une recherche des 10 plus proches voisins. Cette recherche est effectuée selon la distance euclidienne. Ensuite, le spectre attribué est une combinaison linéaire de ces 10 plus proches voisins pondérés par leurs distances.

Approche CNMF-PLS

Pour la deuxième approche, nous proposons d'utiliser la régression des moindres carrés partiels ou *Partial Least Square* (PLS) pour construire un modèle linéaire entre les spectres des pixels et leurs valeurs dans les bandes spectrales de l'image panchromatique. Le modèle est établi à partir du même jeu d'entraînement que celui utilisé par la méthode KNN. C'est-à-dire avec les pixels issus de l'image reconstruite par CNMF et des valeurs contenues dans la pseudo-image panchromatique. Les m spectres de n longueurs d'onde associés à l'image reconstruite CNMF peuvent s'écrire sous forme matricielle avec une matrice $\mathbf{X}_a$ de dimension $\mathbb{R}^{m\times n}$. Les valeurs des m pixels de l'image panchromatique peuvent s'écrire avec une matrice $\mathbf{Y}_a$ de dimension $\mathbb{R}^{m\times 1}$. Les matrices $\mathbf{Y}_a$ et $\mathbf{X}_a$ sont centrées. Soit $\mathbf{b}$ les coefficients de la régression de dimension $\mathbb{R}^{n\times 1}$ et $\mathbf{E}_y$ l'erreur associée, un modèle de régression PLS peut s'écrire de la manière suivante :

$$\mathbf{Y}_a = \mathbf{X}_a\mathbf{b} + \mathbf{E}_y \tag{2.17}$$

Dans notre cas, le nombre de variables latentes du modèle est choisi par validation croisée (ou cross-validation) par bloc formé aléatoirement par un tiers des pixels du

jeu d'entraînement. A partir de ce modèle, il est possible d'attribuer un spectre aux pixels non-appariés en utilisant les coefficients de régression $\mathbf{b}$ du modèle PLS (2.17). Cependant la matrice $\mathbf{b}$ représente un vecteur formé des coefficients de régression. Cette matrice ne fait pas partie des matrices carrées[5]. De ce fait, la matrice $\mathbf{b}$ n'est pas inversible. On utilise alors la matrice pseudo-inverse de Moore-Penrose (Penrose, 1955) : $(\mathbf{bb}')^+$. Soit $\mathbf{Y}_{na}$ les pixels non-appariés contenus dans l'image panchromatique, les spectres de ces pixels sont estimés avec la matrice $\mathbf{X}_{na}$ de la manière suivante :

$$\mathbf{X}_{na} = \mathbf{Y}_{na}\mathbf{b}'(\mathbf{bb}')^+ + \mathbf{E}_x \tag{2.18}$$

Où la matrice $\mathbf{E}_x$ représente l'erreur de prédiction. Les matrices $\mathbf{Y}_{na}$ et $\mathbf{X}_{na}$ sont centrées par rapport au jeu d'entraînement.

2.5.2 Comparaison des deux méthodes proposées

Afin de comparer les deux méthodes proposées en section 2.5.1, nous utilisons l'image panchromatique avec une résolution spatiale fixée à 512x512 et l'image spectrale avec une résolution spatiale fixée à 32x32. Dans une première partie nous faisons varier le taux de recouvrement de l'image spectrale de 100% à 25%. Le taux de recouvrement de l'image spectrale sera défini comme le nombre de pixels appariés sur le nombre total de pixels de l'image panchromatique. Le taux de recouvrement de 100% correspond au cas étudié précédemment où tous les pixels de l'image panchromatique sont appariés. Les indices établis précédemment (CC, SAM et RMSE) sont également utilisés pour comparer l'image de référence avec les images reconstruites par ces deux méthodes.

Fusion d'une image spectrale avec une image panchromatique selon différents taux de recouvrement

Les figures 2.10 présentent l'évolution des valeurs des indices en fonction du taux de recouvrement de l'image spectrale pour les approches CNMF-KNN et CNMF-PLS. L'indice CC (Fig. 2.10a) augmente lorsque le taux de recouvrement augmente pour la méthode CNMF-KNN. En revanche, ce même indice varie très peu selon le taux de recouvrement pour la méthode CNMF-PLS.

Lorsque le taux de recouvrement diminue, la qualité de reconstruction spatiale traduite par CC diminue. Cela est normal, car plus le jeu d'entraînement est petit, plus le risque des pixels du jeu d'entraînement est élevé. Avec le modèle PLS, le recouvrement a peu d'impact sur l'indice CC entre les images. En effet, le nombre d'éléments dans le

5. Les matrices carrées possèdent le même nombre de lignes que de colonnes.

jeu d'entraînement impacte moins le modèle linéaire entre un nombre fixe de spectres et des valeurs de pixels de l'image panchromatique.

En utilisant l'indice SAM (Fig. 2.10b), on constate que plus le taux de recouvrement est grand, plus les valeurs de l'indice SAM sont petites pour les deux approches proposées. De façon générale, les valeurs de l'indice SAM pour la méthode CNMF-KNN sont inférieures à celles obtenues avec la méthode CNMF-PLS. L'écart entre les valeurs de SAM des deux méthodes est d'autant plus grand que le taux de recouvrement est grand.

Pour les deux méthodes, l'augmentation du taux de recouvrement revient à agrandir le jeu d'entraînement et donc à améliorer la qualité spectrale de la reconstruction. Avec la méthode CNMF-KNN, on retrouve les valeurs de SAM des précédentes analyses (Fig. 2.4b). En effet, tous les pixels sont contenus dans le jeu d'entraînement. L'attribution des spectres est directement celle issue de la reconstruction CNMF. En revanche, même avec un taux de recouvrement de 100%, on ne retrouve pas les valeurs de la reconstruction CNMF pour la méthode PLS. En effet, cela vient du choix de variables latentes issues de la validation croisée.

Pour l'indice RMSE (Fig. 2.10c), les valeurs issues du modèle CNMF-PLS diminuent très légèrement en fonction du taux de recouvrement pour passer de 0.105 à 0.10. Pour la méthode CNMF-KNN, l'erreur diminue fortement lorsque le taux de recouvrement augmente.

Le modèle PLS établit une relation linéaire entre les spectres et les valeurs des pixels. Ce modèle est moins sensible au recouvrement. La méthode KNN présente de meilleurs résultats que le modèle PLS pour un taux de recouvrement supérieur à 50%. Cette méthode est plus simple à mettre en oeuvre, à condition d'avoir un jeu d'entraînement suffisant. En effet, si un des pixels est très éloigné du jeu d'entraînement, il sera difficile de lui attribuer un spectre. En revanche le modèle PLS sera plus stable, quel que soit l'éloignement des observations.

Les résultats obtenus via les indices SAM et CC sont des indices moyennés respectivement sur l'ensemble des pixels et sur l'ensemble des longueurs d'onde. Comme cela a été fait précédemment (Fig. 2.5 et 2.6), nous détaillons alors les valeurs de SAM obtenues pour l'ensemble des pixels de l'image et les valeurs de l'indice CC par longueur d'onde. Pour cette étude, nous fixons un taux de recouvrement de 50%.

Les figures 2.11a 2.11b montrent les images formées par les indices SAM obtenus par pixel par pixel pour respectivement la méthode CNMF-KNN et CNMF-PLS. L'image obtenue pour la méthode CNMF-KNN (Fig. 2.11a) montre une pixellisation de certaines régions avec des valeurs d'angle très faibles proches de 1° à 2°. Ces agrégats de pixels correspondent au pixels appariés. Autour de ces régions pixellisées, les autres pixels, lorsqu'ils correspondant à la plante, ont des angles plus élevés de l'ordre de 5 et 10°. Les

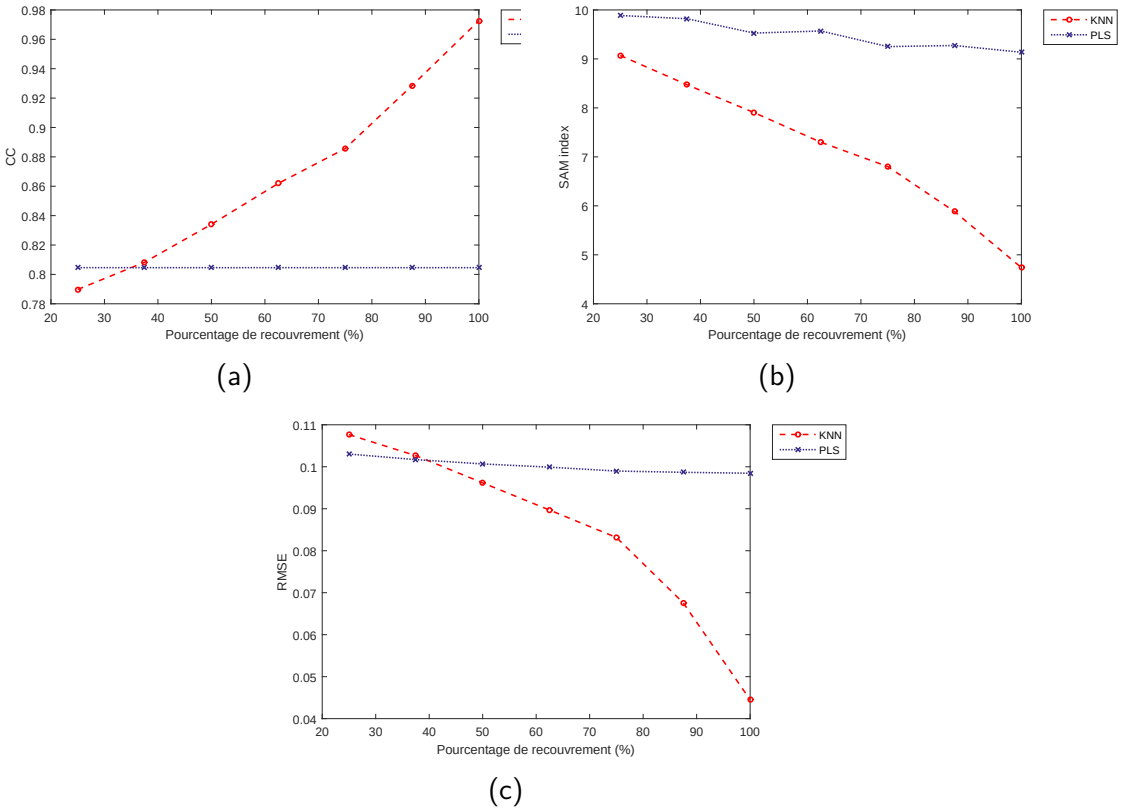


FIGURE 2.10 – Evolution des indices : (a) CC, (b) RMSE, (c), SAM (deg) en fonction du taux de recouvrement pour les approches CNMF-KNN (rouge) et CNMF-PLS (bleu).

pixels correspondant au sol ont des angles beaucoup plus élevés de l'ordre de 20°. Pour la méthode CNMF-PLS, il n'y a pas de discontinuité entre certaines régions de pixels. Pour les pixels correspondant aux feuilles, les valeurs de l'indice SAM sont plus élevées pour la méthode CNMF-PLS que pour la méthode CNMF-KNN. Pour les deux méthodes, les indices SAM sont très élevés pour des pixels de sol avec des valeurs proches de 20°.

La pixellisation de l'image SAM (2.11a) est due à la méthode CNMF-KNN. Les régions des valeurs de SAM proche de 0° correspondent aux pixels contenus dans le jeu d'entraînement. Malgré cela, les autres pixels possèdent des angles beaucoup plus faibles que les angles obtenus via la méthode CNMF-PLS. L'inconvénient majeur de la méthode CNMF-PLS est que les angles obtenus pour les pixels correspondant au jeu d'entraînement sont très élevées. Pour les deux méthodes, il est à noter qu'il est beaucoup plus difficile d'attribuer des spectres aux pixels de sol que de la végétation.

La figure 2.12 montre les valeurs de l'indice CC par bandes spectrales pour chacune des deux méthodes. On remarque que les corrélations dépendent fortement des bandes spectrales. La corrélation entre les images pour des longueurs d'onde inférieures à $700\ \mathrm{nm}$ est beaucoup plus faible que celle pour des longueurs d'onde supérieures à $750\ \mathrm{nm}$. Les valeurs de CC obtenues pour la méthode CNMF-KNN sont supérieures aux valeurs

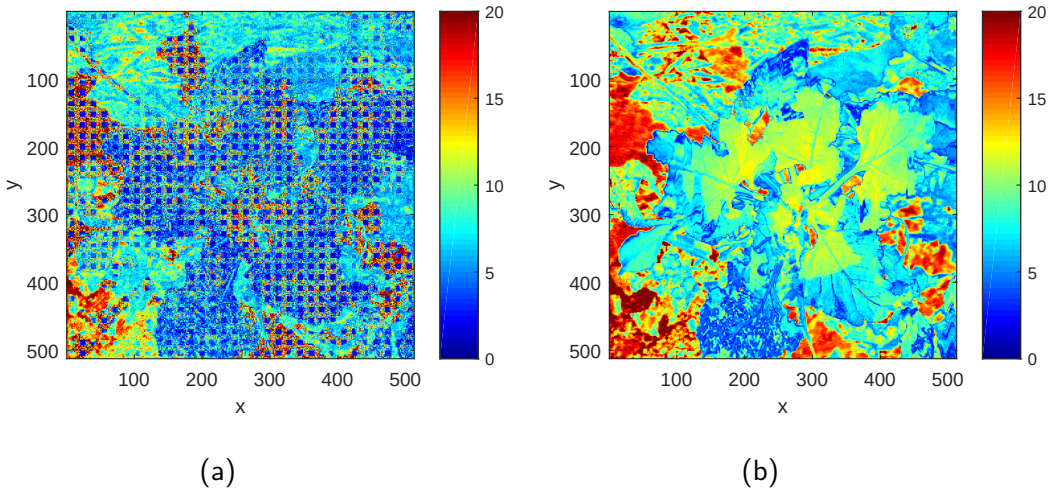


FIGURE 2.11 – Indices SAM pixel à pixel pour les deux méthodes proposées : (a) CNMF-KNN, (b) CNMF-PLS.

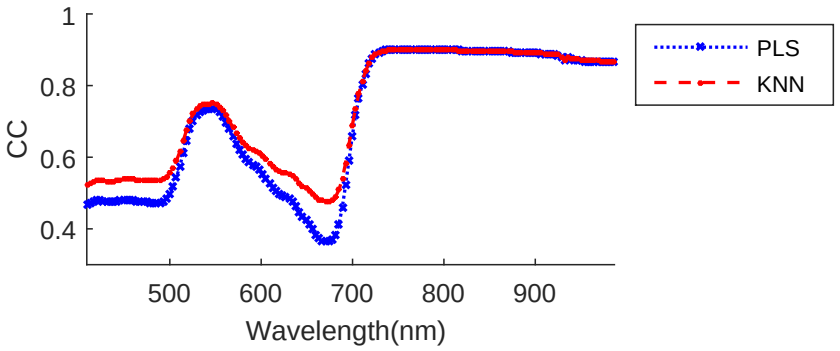


FIGURE 2.12 – Évolution de l'indice CC en fonction de la longueur d'onde pour les deux approches proposées.

obtenues avec la méthode CNMF-PLS, ceci pour l'ensemble des bandes spectrales.

La variation de l'indice CC selon la bande spectrale vient du fait que l'image panchromatique ne possède qu'une seule bande spectrale à une longueur d'onde située à $735\ \mathrm{nm}$. La reconstruction des spectres via cette bande spectrale ne permet de reconstruire qu'une partie du spectre. Pour les bandes spectrales situées aux longueurs inférieures à $700\ \mathrm{nm}$, les valeurs sont largement plus faibles.

Fusion d'une image spectrale avec une image multispectrale selon un taux de recouvrement défini

La question traitée dans cette partie est l'apport d'une image contenant plusieurs bandes spectrales à la place de l'image panchromatique ne contenant qu'une seule bande

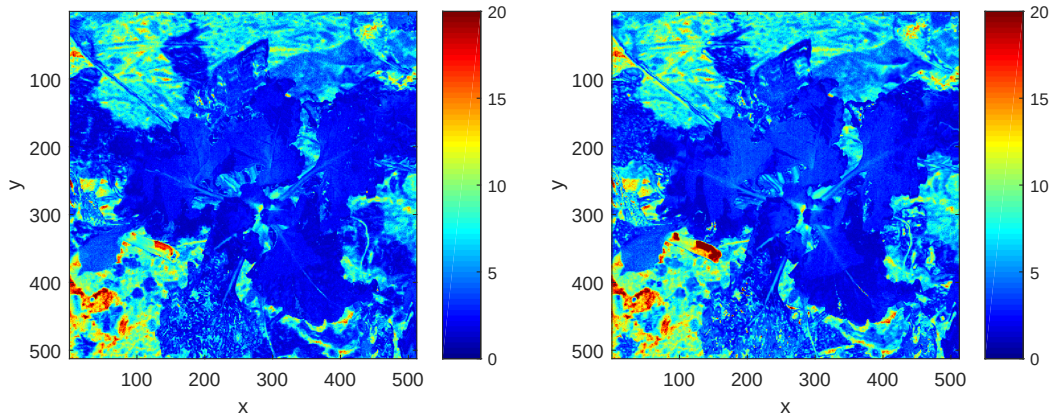


FIGURE 2.13 – Indice SAM pour chaque pixel selon l'utilisation les méthodes : (a) KNN, (b) PLS.

spectrale. Comme précédemment, nous fixons un taux de recouvrement à 50% pour comparer les méthodes CNMF-KNN et CNMF-PLS dans ce cas. L'image multispectrale remplace alors l'image panchromatique précédente. L'image multispectrale contient alors trois bandes correspondant ici à de l'imagerie classique couleur ou RGB. Les bandes spectrales sont situées aux longueurs d'onde suivantes : $605\,\mathrm{nm}$ (R), $550\,\mathrm{nm}$ (G) et $450\,\mathrm{nm}$ (B).

La table 2.3 montre les valeurs obtenues pour les indices CC, SAM et RMSE pour les deux méthodes testées en utilisant une image multispectrale. Pour la méthode CNMF-KNN, les valeurs sont de 0.98, 4.56° et de 0.09 pour respectivement les indices CC, SAM et RMSE. Pour la méthode CNMF-PLS, les valeurs sont de 0.96, 5.02° et de 0.11 pour respectivement les indices CC, SAM et RMSE.

Pour un taux de recouvrement identique (50%), les résultats issus de la fusion avec l'image couleur sont meilleurs que ceux obtenus avec l'image panchromatique (Fig. 2.10). Les indices CC sont plus proches de 1, les indices SAM sont réduits d'environ 3° et les RMSE restent équivalents. En fusionnant avec l'image multispectrale, l'écart entre les deux méthodes est plus faible. Toutefois la méthode CNMF-KNN fournit des meilleurs résultats sur l'ensemble de l'image étudiée.

TABLE 2.3 – Valeurs des indices CC, SAM et RMSE pour les deux approches testées en utilisante une image multispectrale.

Méthode	CC	SAM	RMSE
KNN	0.98	4.56	0.09
PLS	0.96	5.02	0.11

Les figures 2.13 montrent les valeurs de l'indice SAM obtenues pixel par pixel. Pour

la méthode CNMF-KNN (Fig. 2.13a), les pixels liés à la végétation sont proches de 0°. Certains pixels de sols et du tuteur sont en revanche très différents de l'image de référence avec des angles de 17°. Pour la méthode CNMF-PLS (Fig. 2.13b), les spectres des pixels de végétation sont également similaires à l'image de référence avec des angles proches de 0°. Des angles plus importants apparaissent pour les pixels liés au tuteur et au sol avec des valeurs d'angle proches de 20°. Quelle que soit la méthode utilisée, ce sont les mêmes régions où les pixels sont relativement différents de l'image de référence.

On retrouve des valeurs d'angles très faibles pour les deux méthodes. Des légères différences sont visibles pour les mêmes pixels de l'image, à savoir les pixels liés au sol et au tuteur. De même, pour les deux méthodes, les spectres de végétation sont très similaires à ceux de l'image de référence. Cette vérification permet de vérifier quels sont les éléments de la scène qui sont bien reconstruits. On peut supposer que le jeu d'entraînement contient beaucoup de pixels de végétation. En effet, on retrouve beaucoup plus de pixels de végétation dans l'image que de pixels de sol.

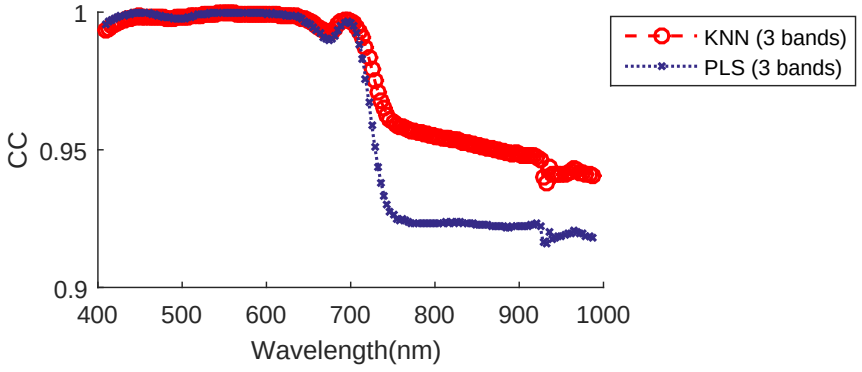


FIGURE 2.14 – Indice CC par bandes spectrales des méthodes proposées après la fusion entre l'image hyperspectrale.

La figure 2.14 montre les indices CC par longueur d'onde des deux méthodes issues de la fusion de l'image hyperspectrale et de l'image multispectrale. Les valeurs obtenues pour les deux méthodes sont très proches de 1 lorsque les longueurs d'onde sont inférieures à $700\ \mathrm{nm}$ et diminuent pour des longueurs d'onde supérieures. Des valeurs minimales de 0.94 et de 0.92 sont atteintes pour respectivement la méthode CNMF-KNN et CNMF-PLS.

L'usage d'une image multispectrale avec 3 bandes situées dans la région visible permet de bien reconstruire l'ensemble des bandes spectrales situées entre $400\ \mathrm{nm}$ à $740\ \mathrm{nm}$. Après $750\ \mathrm{nm}$, la qualité est inférieure. Néanmoins les valeurs de CC obtenues sont proches de celles obtenues avec l'image panchromatique (Fig. 2.12). L'image multispectrale étant formée de bandes spectrales situées dans le visible, l'image résultante dans

ce domaine spectral est alors bien reconstruite.

2.5.3 Conclusion partielle

Lorsque l'on combine plusieurs capteurs, des pixels peuvent être non-appariés. Cela peut apparaître pour différentes raisons comme une différence de champ de vue [6] ou d'une différence de temps d'intégration entre les deux imageurs.

Deux méthodes permettant d'attribuer des spectres aux pixels non-appariés ont alors été proposées. Il s'agit des méthodes CNMF-KNN et CNMF-PLS. Ces méthodes ont été testées dans le cadre d'une fusion d'images panchromatique et spectrale en faisant varier le taux de recouvrement.

La méthode CNMF-KNN est très sensible au taux de recouvrement alors que les résultats de la méthode CNMF-PLS varient faiblement en fonction de celui-ci. En effet, les performances de la méthode KNN sont associées à la dimension du jeu d'entrainement. Plus celle-ci sera faible, plus il sera difficile de faire correspondre les observations. L'ensemble des résultats obtenus ne sont pas concluants pour l'ensemble des bandes spectrales. Dans la partie visible, les corrélations images reconstruites/références sont très faibles. Les deux méthodes ont néanmoins montré quelques défauts : à savoir, une mauvaise prédiction des pixels de végétation par la méthode CNMF-PLS et une pixellisation de la méthode CNMF-KNN entre jeu d'entraînement et pixels à estimer.

Remplacer l'image panchromatique par une image multispectrale de type caméra traditionnelle (RGB) améliore grandement la qualité de reconstruction spectrale et spatiale via l'utilisation des deux méthodes proposées. On a pu aussi démontrer le potentiel de ces méthodes pour combiner des capteurs ayant des recouvrements différents.

2.6 Conclusion du chapitre

Dans ce chapitre, un état de l'art des méthodes existantes de pan-sharpening a été réalisé. Une comparaison des principales méthodes appartenant aux 3 classes de méthodes identifiées a été réalisée à l'aide d'une approche par simulation utilisant une image hyperspectrale de référence. La classe de méthodes basée sur le démélange spectral s'est montrée particulièrement performante avec notamment la méthode CNMF. De plus, cette méthode requiert un très faible nombre de spectres par rapport aux nombres de pixels d'une image panchromatique. Une résolution spatiale minimale doit tout de même être fixée pour l'image spectrale.

A partir de la méthode CNMF, nous avons proposé deux méthodes permettant de

6. Field Of View : FOV

traiter spécifiquement le problème de manque de recouvrement. Le manque de recouvrement a été défini comme des pixels non-appariés, c'est à dire à des pixels de l'image panchromatique non associés à des pixels de l'image spectrale. Nous avons vu que la variation du taux de recouvrement peut jouer un rôle sur la qualité de reconstruction.

Un taux de recouvrement faible présente un avantage, à savoir obtenir des détails de certains pixels et donc d'accéder aux spectres d'objets de petites dimensions (nervures de feuilles par exemple). En revanche, cela entraîne une augmentation du nombre des pixels non-appariés et donc une augmentation d'attributions de spectres. Il faut alors trouver un compromis entre le taux de recouvrement et la résolution spatiale de l'image spectrale (et donc du nombre de spectres dans le jeu d'entraînement).

Attribuer un spectre à un pixel ne contenant qu'une bande spectrale est difficile pour un taux de recouvrement de 50%. Ici l'apport de plusieurs bandes spectrales à l'aide de l'imagerie multispectrale permet d'améliorer les résultats de la reconstruction. Ces résultats permettent d'envisager de concevoir un système combinant l'imagerie multispectrale et des mesures spectrales à des points spécifiques.

Les résultats peuvent être encore améliorés en identifiant précisément les bandes spectrales permettant une meilleure reconstruction de l'image. Cette identification permettrait de choisir la caméra multispectrale la plus adaptée pour le couplage de capteurs. Des solutions d'imagerie basées sur l'interférométrie de type Fabry-Pérot permettent d'ailleurs d'ajuster les bandes spectrales (Honkavaara et al., 2017; Näsi et al., 2016).

Les méthodes proposées, associées à un dispositif combinant capteur faible résolution spatiale/forte résolution spectrale et capteur forte résolution spatiale/faible résolution spectrale doivent être validées dans des conditions réelles d'utilisation. Cette validation doit s'effectuer en comparant ce dispositif avec un capteur hyperspectral.

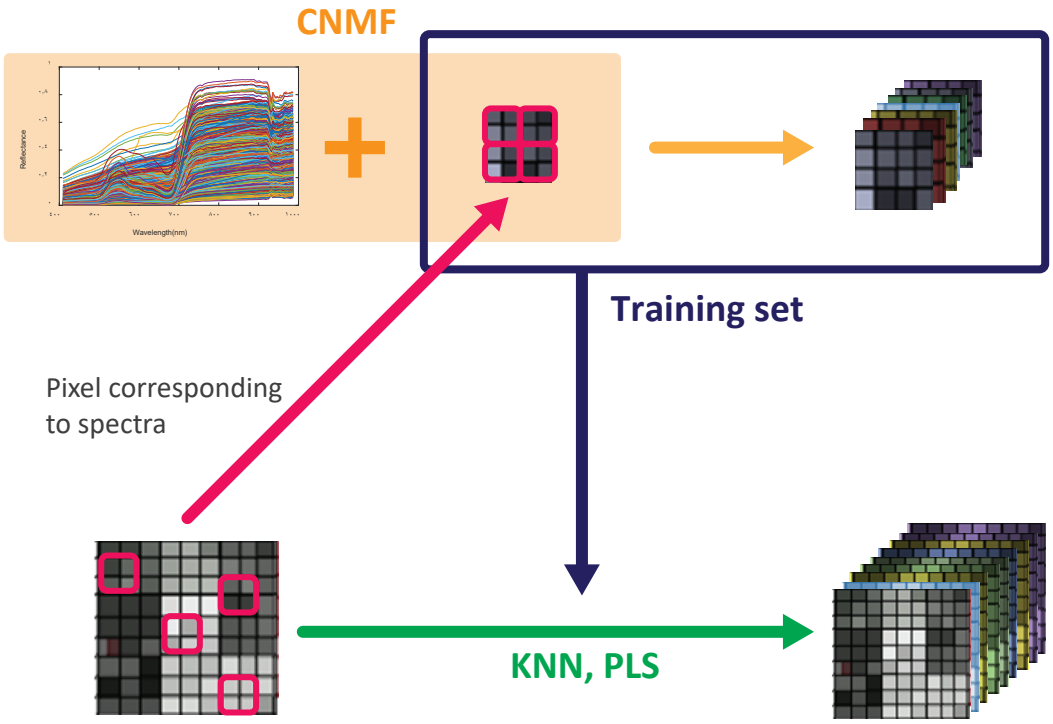


FIGURE 2.15 – Synthèse du chapitre 2 : Méthode proposée pour reconstruire une image hyperspectrale à partir d'une image panchromatique ou multispectrale et des informations spectrales sur cette image.

Bibliographie

F. Cliche. Integration of the SPOT panchromatic channel into its multispectral mode for image sharpness enhancement. *PE & RS*, 51(3) :311–316, 1985.

R. Welch. Cartographic potential of SPOT image data. *Photogrammetric Engineering and Remote Sensing(ISSN 0099-1112)*, 51 :1085–1091, 1985.

Pat S. Chavez Jr. Digital processing techniques for image mapping with Landsat TM and SPOT simulator data. In *Proceedings of the Eighteenth International Symposium on Remote Sensing of Environment.*, volume 1, pages 101–116, 1984.

Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. A Critical Comparison Among Pansharpening Algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5) :2565–2586, May 2015. ISSN 0196-2892, 1558-0644. doi : 10.1109/TGRS.2014.2361734. URL http://ieeexplore.ieee.org/document/6998089/.

Rosa Lasaponara and Nicola Masini. Pan-sharpening techniques to enhance archaeological marks : an overview. In *Satellite Remote Sensing*, pages 87–109. Springer, 2012.

I. Ozdemir. Estimating stem volume by tree crown area and tree shadow area extracted from pan-sharpened Quickbird imagery in open Crimean juniper forests. *International Journal of Remote Sensing*, 29(19) :5643–5655, October 2008. ISSN 0143-1161. doi : 10.1080/01431160802082155. URL https://doi.org/10.1080/01431160802082155.

Khan Rahaman, Quazi Hassan, and M. Ahmed. Pan-sharpening of Landsat-8 images and its application in calculating vegetation greenness and canopy water contents. *ISPRS International Journal of Geo-Information*, 6(6) :168, 2017.

M. P. Labus, G. A. Nielsen, R. L. Lawrence, R. Engel, and D. S. Long. Wheat yield estimates using multi-temporal NDVI satellite imagery. *International Journal of Remote Sensing*, 23(20) :4169–4180, January 2002. ISSN 0143-1161, 1366-5901. doi : 10.1080/01431160110107653. URL https://www.tandfonline.com/doi/full/10.1080/01431160110107653.

Roberto Colombo, Dario Bellingeri, Dante Fasolini, and Carlo M. Marino. Retrieval of leaf area index in different vegetation types using high resolution satellite data. *Remote Sensing of Environment*, 86(1) :120–131, June 2003. ISSN 0034-4257. doi : 10.1016/S0034-4257(03)00094-4. URL http://www.sciencedirect.com/science/article/pii/S0034425703000944.

Caroline M. Gevaert, Juha Suomalainen, Jing Tang, and Lammert Kooistra. Generation of Spectral–Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6) :3140–3146, June 2015. ISSN 1939-1404, 2151-1535. doi : 10.1109/JSTARS.2015.2406339. URL http://ieeexplore.ieee.org/document/7058421/.

Agnieszka Jenerowicz and Malgorzata Woroszkiewicz. The pan-sharpening of satellite and UAV imagery for agricultural applications. page 99981S, October 2016. doi : 10.1117/12.2241645. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2241645.

Philippe Blanc, Lucien Wald, and Thierry Ranchin. Importance and effect of co-registration quality in an example of "pixel to pixel" fusion process. In *2nd Interna-*

tional Conference "Fusion of Earth Data : merging point measurements, raster maps and remotely sensed images", pages 67–74. SEE/URISCA, Nice, France, 1998.

Laetitia Loncan, Luis B. de Almeida, José M. Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A. Licciardi, Miguel Simoes, and others. Hyperspectral pansharpening : a review. *IEEE Geoscience and remote sensing magazine*, 3(3) :27–46, 2015.

Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2) :528–537, 2012. URL http://ieeexplore.ieee.org/abstract/document/5982386/.

Miguel Simões, José Bioucas-Dias, Luis B. Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6) :3373–3388, 2015. URL http://ieeexplore.ieee.org/abstract/document/7000523/.

Qi Wei, Nicolas Dobigeon, and Jean-Yves Tourneret. Bayesian fusion of multi-band images. *IEEE Journal of Selected Topics in Signal Processing*, 9(6) :1117–1127, 2015a.

Qi Wei, José Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7) :3658–3668, 2015b.

W. Joseph Carper. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm. Eng. Remote Sens.*, 56(4) :457–467, 1990. URL http://ci.nii.ac.jp/naid/80005825315/.

Te-Ming Tu, Shun-Chi Su, Hsuen-Chyun Shyu, and Ping S. Huang. A new look at IHS-like image fusion methods. *Information fusion*, 2(3) :177–186, 2001. URL http://www.sciencedirect.com/science/article/pii/S1566253501000367.

Pats Chavez, Stuart C. Sides, and Jeffrey A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data- Landsat TM and SPOT panchromatic. *Photogrammetric Engineering and remote sensing*, 57(3) :295–303, 1991. URL http://www.cis.rit.edu/~cnspci/references/dip/fusion/chavez1991.pdf.

P. S. Kwarteng and A. Y. Chavez. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens*, 55 :339–348, 1989.

Alan R. Gillespie, Anne B. Kahle, and Richard E. Walker. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3) :343–365, 1987. URL http://www.sciencedirect.com/science/article/pii/0034425787900885.

Craig A. Laben and Bernard V. Brower. *Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening*. Google Patents, January 2000. URL https://www.google.com/patents/US6011875.

Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of MS $+ $ Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10) :3230–3239, 2007.

J. G. Liu. Smoothing filter-based intensity modulation : A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21 (18) :3461–3472, 2000. URL http://www.tandfonline.com/doi/abs/10.1080/014311600750037499.

Luciano Alparone, Lucien Wald, Jocelyn Chanussot, Claire Thomas, Paolo Gamba, and Lori Mann Bruce. Comparison of pansharpening algorithms : Outcome of the 2006 GRS-S data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10) :3012–3021, 2007.

B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva. An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas. In *2003 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, pages 90–94, May 2003. doi : 10.1109/DFUA.2003.1219964. ISSN : null.

Wenzhi Liao, Xin Huang, Frieke Van Coillie, Sidharta Gautama, Aleksandra Pizurica, Wilfried Philips, Hui Liu, Tingting Zhu, Michal Shimoni, Gabriele Moser, and Devis Tuia. Processing of Multiresolution Thermal Hyperspectral and Digital Color Data : Outcome of the 2014 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6) :2984–2996, June 2015. ISSN 1939-1404, 2151-1535. doi : 10.1109/JSTARS.2015.2420582. URL http://ieeexplore.ieee.org/document/7109121/.

Fatima Zohra Benhalouche, Moussa Sofiane Karoui, Yannick Deville, and Abdelaziz Ouamri. Hyperspectral data multi-sharpening based on linear-quadratic nonnegative matrix factorization. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3290–3293. IEEE, 2015.

Cyril Ruckebusch, editor. *Resolving spectral mixtures : with applications from ultrafast time-resolved spectroscopy to super-resolution imaging*. Number Volume 30 in Data handling in science and technology. Elsevier, Amsterdam Heidelberg, first edition edition, 2016. ISBN 978-0-444-63638-6.

Nicolas Dobigeon, Yoann Altmann, N. Brun, and S. Moussaoui. Linear and nonlinear unmixing in hyperspectral imaging. In *Data Handling in Science and Technology*, volume 30, pages 185–224. Elsevier, 2016.

Xuesong Liu, Wei Xia, Bin Wang, and Liming Zhang. An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2) :757–772, 2010.

Miguel Simoes, José Bioucas-Dias, Luis B. Almeida, and Jocelyn Chanussot. Hyperspectral image superresolution : An edge-preserving convex formulation. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 4166–4170. IEEE, 2014. URL http://ieeexplore.ieee.org/abstract/document/7025846/.

José MP Nascimento and José MB Dias. Vertex component analysis : A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4) :898–910, 2005. URL http://ieeexplore.ieee.org/abstract/document/1411995/.

Xavier Bresson and Tony F. Chan. Fast dual minimization of the vectorial total variation norm and applications to color image processing. *Inverse problems and imaging*, 2(4) : 455–484, 2008.

Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions : assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6) :691–699, 1997. URL https://hal.archives-ouvertes.fr/hal-00365304/.

F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3) :145–163, 1993. URL http://www.sciencedirect.com/science/article/pii/003442579390013N.

Roberta H. Yuhas, Alexander FH Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. 1992. URL https://ntrs.nasa.gov/search.jsp?R=19940012238.

Muhammad A. Sharaf, Deborah L. Illman, and Bruce R. Kowalski. *Chemometrics*. John Wiley & Sons, March 1986. ISBN 978-0-471-83106-8. Google-Books-ID : JQQC8ISIXB8C.

Katherine A. Bakeev, editor. *Process analytical technology : spectroscopic tools and implemented strategies for the chemical and pharmaceutical industries*. Wiley, Chichester, West Sussex, 2nd ed edition, 2010. ISBN 978-0-470-72207-7. OCLC : ocn473478595.

R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3) :406–413, July 1955. ISSN 0305-0041, 1469-8064. doi : 10.1017/S0305004100030401. URL https://www.cambridge.org/core/product/identifier/S0305004100030401/type/journal_article.

Eija Honkavaara, Tomi Rosnell, Raquel Oliveira, and Antonio Tommaselli. Band registration of tuneable frame format hyperspectral UAV imagers in complex scenes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134 :96–109, December 2017. ISSN 0924-2716. doi : 10.1016/j.isprsjprs.2017.10.014. URL http://www.sciencedirect.com/science/article/pii/S0924271616305378.

Roope Näsi, E. Honkavaara, Sakari Tuominen, H. Saari, Ilkka Pölönen, T. Hakala, N. Viljanen, J. Soukkamäki, I. Näkki, and H. Ojanen. UAS based tree species identification using the novel FPI based hyperspectral cameras in visible, NIR and SWIR spectral ranges. *International archives of the photogrammetry, remote sensing and spatial information sciences ; Volume XLI-B1*, 2016.

Proposition d'une méthode de correction des erreurs de répétabilité pour l'analyse de variance

Chapitre 3

Sommaire

3.1 Introduction **61**

3.2 Prérequis et Définitions **62**

3.2.1 Plan d'expérience et analyse de variance 62

3.2.2 Espace colonne et ligne ou Column and Row space 63

3.2.3 Orthogonalisation 64

3.3 Théorie de REP-ASCA **65**

3.3.1 ASCA 65

3.3.2 Erreur de répétabilité 66

3.3.3 REP-ASCA 67

3.4 Test de REP-ASCA par simulation **68**

3.4.1 Préparation de données simulées 69

3.4.2 Résultats et discussion 69

3.4.3 Conclusion de la simulation 71

3.5 Conclusion du chapitre **72**

3.1 Introduction

L'objectif d'une expérience est d'étudier les variables dites explicatives, communément appelées facteurs, qui sont susceptibles d'avoir une influence sur une réponse d'un système étudié. L'ensemble des réponses (ou observations) du système étudié constitue ainsi un jeu de données. Des méthodes d'analyse de résultats issus d'une expérience appelées analyses de variance permettent de tester les influences des facteurs sur la variabilité de ce jeu de données.

Cependant, lorsque l'on répète plusieurs fois une mesure sur un échantillon dans les mêmes conditions, on constate que ces mesures peuvent varier. Cela s'interprète comme un manque de répétabilité entre des mesures réalisées sur un même individu. Ces erreurs de répétabilité contiennent des effets non désirés qui peuvent entacher alors les conclusions d'une telle analyse de variance.

Dans le cas de données univariées, certaines erreurs peuvent être décrites comme des facteurs aléatoires. Ces facteurs aléatoires sont pris en compte dans des modèles de type Generalized Linear Mixed Models (GLMMs) (Bolker et al., 2009). Dans le cas de données multivariées, les modèles de type GLMMs sont inappropriés dans la majeure partie des cas, spécialement avec les données spectrales (Anderson and Braak, 2003).

Nous proposons donc dans ce chapitre une méthode appelée REP-ASCA permettant de réduire les erreurs de répétabilités de jeu de données multivariées destiné à l'analyse de variance. Dans un premier temps, nous donnerons quelques définitions sur l'analyse de variance et sur les traitements multivariés usuels. Puis, nous décrirons la méthode proposée REP-ASCA.

3.2 Prérequis et Définitions

3.2.1 Plan d'expérience et analyse de variance

Historiquement, la méthodologie basée sur le plan d'expériences et l'analyse de variance proposée par Fisher (Fisher, 1937) a été à l'origine d'un ensemble d'études scientifiques couvrant plusieurs disciplines (Stanley, 1962). Le plan d'expérience aide l'expérimentateur à organiser de façon optimale ses expériences pour séparer au mieux les effets liés aux différents facteurs (Fisher, 1937; Oehlert, 2000). De manière générale, l'analyse de variance peut se définir comme une classe de méthodes d'analyse statistique des résultats issues d'une expérimentation. Analyse de variance ou ANOVA fait référence également à la première méthode développée pour l'étude de données univariées (Fisher, 2006), c'est-à-dire pour l'étude d'une seule variable de réponse.

Les variables explicatives contrôlées par l'expérimentateur sont appelées **traitement** et les variables explicatives observées sont appelées **facteurs**. Les facteurs sont catégoriels mais peuvent provenir de variables continues, discrètes, ordonnables ou booléennes. Au moment de l'analyse, il convient alors d'établir les valeurs que peut prendre le facteur. Ces valeurs sont appelées **niveaux** ou modalités du facteur. Par exemple, si l'expérience prévoit une mesure de longueur, le facteur "longueur" se base sur une variable continue ordonnable. Des niveaux peuvent être définis comme par exemple : un premier niveau composé des observations où les valeurs de longueurs sont supérieures ou égales à 5

mètres. Et un deuxième niveau composé des observations où les longueurs sont inférieures à 5 mètres. Cet exemple montre qu'il est possible d'obtenir un facteur, ou variable catégorielle à partir d'une variable continue.

Un modèle d'analyse de variance est établi en identifiant l'ensemble des facteurs. Ces facteurs sont dits statistiquement significatifs dès lors qu'ils agissent sur la réponse d'un ensemble d'observations. On dit également que les facteurs ont un effet sur la variance. Pour évaluer l'effet d'un facteur sur la réponse, un test est réalisé. Ce test fournit une p-value qui représente la probabilité qu'un facteur soit négligeable avec l'hypothèse formulée. L'hypothèse formulée correspond au modèle d'analyse de variance sous-jacent.

Lorsque l'on souhaite étudier le comportement de plusieurs variables en même temps, on utilise l'analyse de variance multiple (MANOVA). L'ANOVA ou la MANOVA repose sur le test de Fisher, il faut donc respecter certaines conditions et vérifier qu'elles sont respectées. Les conditions portent sur l'équivalence des variances appelée aussi homoscédasticité, la normalité des distributions des valeurs et aussi l'indépendance des observations. En revanche, l'ensemble de ces conditions ne sont pas respectées pour certains types de données multivariées et notamment les données spectrales (Anderson and Braak, 2003).

Pour des données multivariées, des méthodes issues de la chimiométrie permettent de tirer parti du plan d'expérience (Brereton et al., 2017). La méthode la plus utilisée est la méthode Analysis of Variance – Simultaneous Component Analysis (ASCA) (Smilde et al., 2005).

3.2.2 Espace colonne et ligne ou Column and Row space

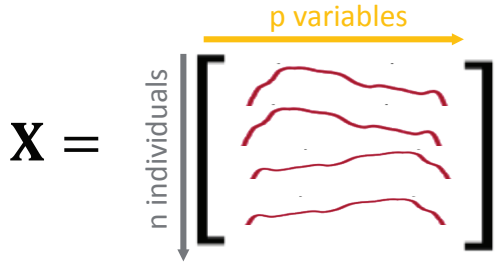


FIGURE 3.1 – Représentation matricelle d'une collection d'observations.

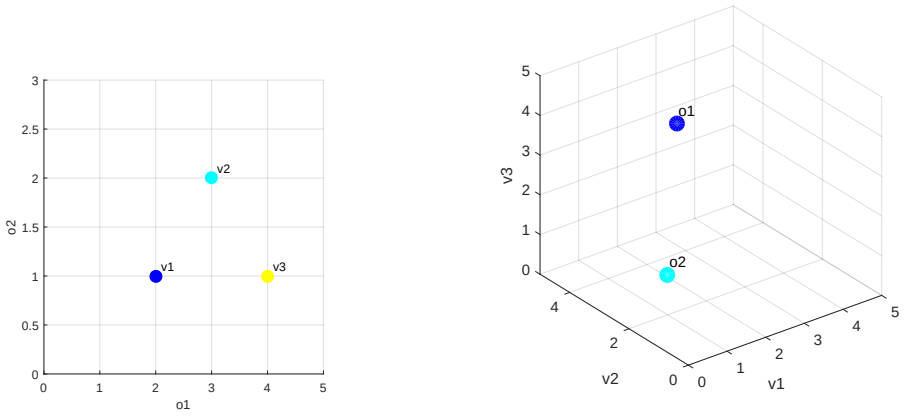
Les données multivariées peuvent être représentées avec une matrice $\mathbf{X}$ de taille (N, P) où les N-lignes représentent les mesures/observations réalisées sur P-variables (Fig. 3.1). Cette matrice peut être représentée dans deux espaces : le row-space et le column-

space. Le row-space est le sous-espace de dimensions $\mathbb{R}^P$ engendré par les lignes de la matrice $\mathbf{X}$. Le column-space est le sous-espace de dimensions $\mathbb{R}^N$ engendré par les colonnes de la matrice $\mathbf{X}$.

Soit une matrice $\mathbf{X}$ définie de la manière suivante :

$$\mathbf{X} = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 2 & 1 \end{pmatrix} \tag{3.1}$$

Où les deux lignes [2 3 4] et [1 2 1] correspondent respectivement à o1 et o2. Les variables sont définies par v1=[2 1], v2=[3 2] et v3=[4 1]. Alors le column-space correspond à l'espace formé par les observations o1 et o2 (Fig. 3.2a). La dimension de cet espace dépend du nombre d'observations. Il varie donc en fonction des expérimentations. Le row-space est l'espace formé par les variables v1, v2 et v3 (Fig. 3.2b). La dimension de cet espace ne varie pas en fonction du nombre d'observations.



(a) (b)

FIGURE 3.2 – Représentation des espaces : (a) formé par les observations (ou column-space) (b) formé par les variables (ou row-space).

3.2.3 Orthogonalisation

Des perturbations peuvent apparaître dans un signal mesuré et peuvent introduire un biais dans l'ensemble des mesures. Une méthode appelée External Parameter Orthogonalisation (Roger et al., 2003) a été développée pour retirer le sous-espace $\mathbf{A}$ lié aux perturbations. Le sous-espace $\mathbf{A}$ est de dimension $\mathbb{R}^{P\times Q}$ où Q est le nombre de composantes permettant de décrire les perturbations. Le sous-espace de la matrice $\mathbf{X}$ lié aux perturbations de la matrice $\mathbf{A}$ peut être identifié par la projection orthogonale $\mathcal{P}_{\mathcal{A}}$:

$$\mathcal{P}_{\mathcal{A}} = \mathbf{A}(\mathbf{A}^t\mathbf{A})\mathbf{A}^t \tag{3.2}$$

Soit $\mathbf{I}$ la matrice identité, la matrice $\mathbf{X}$ peut alors être orthogonalisée selon l'équation suivante :

$$\mathbf{X}_{\perp} = \mathbf{X}(\mathbf{I} - \mathcal{P}_{\mathcal{A}}) \tag{3.3}$$

Ici, la perturbation est traduite par un signal de dimension $\mathbb{R}^{P \times Q}$. La projection orthogonale s'effectue alors dans le row-space. Mais il est possible que l'espace des perturbations soit défini comme un sous-espace lié aux observations. Dans ce cas la dimension du sous-espace serait de dimension $\mathbb{R}^{Q \times N}$.

3.3 Théorie de REP-ASCA

3.3.1 ASCA

L'analyse de variance établit un modèle de décomposition de la matrice $\mathbf{X}$ en différentes sources de variances. Cette décomposition permet d'étudier les relations existantes entre les facteurs et les variables mesurées. L'ASCA est ainsi divisée en trois étapes : La première consiste à décomposer la matrice d'observation $\mathbf{X}$ en matrices associées aux effets étudiés et à leurs interactions. Ainsi par exemple, dans un plan d'expérience impliquant deux facteurs étudiés A et B, la matrice d'observations $\mathbf{X}$ se décompose selon l'approche ASCA de la manière suivante :

$$\mathbf{X} = \mu + \mathbf{X}_A + \mathbf{X}_B + \mathbf{X}_{AB} + \mathbf{E} \tag{3.4}$$

Avec μ la moyenne de l'ensemble des observations, $\mathbf{X}_A$ et $\mathbf{X}_B$ les termes d'effets principaux des facteurs A et B, $\mathbf{X}_{AB}$, le terme d'interaction entre A et B et E les résidus. Les matrices $\mathbf{X}_A$, $\mathbf{X}_B$ et $\mathbf{X}_{AB}$ contiennent les spectres moyens de chacune des modalités pour respectivement les facteurs A, B et AB. Par exemple, toutes les lignes de $\mathbf{X}_A$ correspondant aux échantillons de la modalité 1 du facteur A contiennent le même spectre. Ce spectre est le spectre moyen de la modalité 1 du facteur A. Dans la pratique, les variances de chacun des facteurs sont remplacées par la somme des carrés (SSQ) des éléments de la matrice associée. Par exemple, la variance portée par la matrice $\mathbf{X}_A$ est remplacée par (De Luca et al., 2016; Smilde et al., 2008; Vis et al., 2007) :

$$\mathrm{SSQ}(\mathbf{X}_A) = \|\mathbf{X}_A\|^2 \tag{3.5}$$

Où $\|\mathbf{X}_A\|^2$ correspond à la norme carrée de Frobenius de $\mathbf{X}_A$. Dans le cas d'un plan d'expérience complet et équilibré, les facteurs A, B et AB sont indépendants entre eux (Shaw and Mitchell-Olds, 1993). Dans ce cas précis, ces facteurs sont dits orthogonaux

(Smilde et al., 2008), car les matrices $\mathbf{X}_A$, $\mathbf{X}_B$ et $\mathbf{X}_{AB}$ sont orthogonales deux à deux dans le column-space. Ainsi on a : $SSQ(\mathbf{X}) = SSQ(\mathbf{X}_A) + SSQ(\mathbf{X}_B) + SSQ(\mathbf{X}_{AB}) + SSQ(\mathbf{E})$. Le terme $SSQ(\mathbf{E})$ contient la variance intra-modalité et les autres termes la variance inter-modalité.

La deuxième étape de l'ASCA est de décrire l'effet de chacun des facteurs. Tout d'abord, pour savoir si les facteurs ont un effet statistiquement significatif sur la variance, un test de permutation est réalisé. Ce test permet d'estimer une p-value pour des données ne respectant pas forcément les conditions de normalités (Anderson and Braak, 2003). Pour un facteur i et sa matrice associée $\mathbf{X}_i$, la p-value se calcule selon l'équation suivante :

$$\text{p-value}(\mathbf{X}_i) = \frac{\text{nbr}(\text{SSQ}(\mathbf{X}_{i,perm} \geq \text{SSQ}(X_i))}{k} \tag{3.6}$$

Où k est le nombre de permutations réalisées et $\mathbf{X}_{i,perm}$ la matrice obtenue après une permutation aléatoire des lignes. La p-value est donc calculée en recensant le nombre de cas où la variance du facteur étudié est inférieure aux variances issues des permutations. En d'autres termes, la variance du facteur étudié est comparée à la distribution de valeurs de variances de ce même facteur formées par les permutations. Pour la dernière étape, une Simultaneous Component Analysis (SCA) (Kiers, 1990; Van Deun et al., 2009) est réalisée sur l'ensemble des termes des matrices présentes dans l'équation (ASCA). Ainsi, pour un facteur donné i, la décomposition de la matrice correspondante $\mathbf{X}_i$ s'écrit :

$$\mathbf{X}_i = \mathbf{T}_i \mathbf{P}_i^t \tag{3.7}$$

Où $\mathbf{T}_i$ et $\mathbf{P}_i$ correspondent respectivement aux scores et aux loadings des composantes principales. Cette analyse permet de réduire l'espace de représentation. Ces loadings permettent de définir un sous-espace lié au terme $\mathbf{X}_i$ et ainsi de mettre en évidence les directions spectrales liées aux facteurs étudiés. Les scores permettent de classer les observations.

3.3.2 Erreur de répétabilité

Lorsqu'on effectue plusieurs mesures sur un même échantillon, les différentes mesures ne sont généralement pas identiques. Cette différence est due à l'erreur de répétabilité. Lorsqu'on effectue des mesures sur des échantillons différents, les différences constatées contiennent à la fois l'erreur de répétabilité et les variations dues aux différences entre les échantillons. Si l'erreur de répétabilité est importante, la variance intra modalité augmente. Les variances portées par les facteurs deviennent plus faibles par rapport à la variance totale. Ceci qui a pour tendance de diminuer la significativité des facteurs.

L'erreur de répétabilité peut être due à la variation des conditions d'observations

(e.g. l'angle de vue, la prise de la référence) ou de l'échantillon mesuré (e.g. la taille des particules, rugosité, texture physique, ...). De plus, la composition chimique non-uniforme modifie la diffusion de la lumière dans le milieu. Par exemple, pour un même échantillon, une variation de la teneur en huile, eau, en fibres, en protéines et en matières minérales change la nature du milieu modifiant la diffusion de la lumière. Mesurer des échantillons non-uniformes ajoute une difficulté majeure par rapport à des échantillons de poudres homogènes (Williams and Norris, 1987). L'ensemble de ces éléments peuvent entraîner des déformations du spectre, telles que l'apparition de lignes de bases, d'effets multiplicatifs, de shifts en longueurs d'onde, etc.

L'erreur de répétabilité est généralement le résultat de phénomènes aléatoires d'espérance nulle. En conséquence, dans ASCA, l'erreur de répétabilité est réduite car le spectre d'un niveau d'une modalité i du facteur j est la moyenne de tous les spectres de cette modalité. En effet, moyenner permet d'approximer l'espérance. Malgré cela, lorsque cette réduction n'est pas suffisante, la solution couramment employée est de répéter les mesures réalisées sur chaque échantillon. Cela a pour conséquence d'augmenter considérablement le nombre d'acquisitions. De plus, il est impossible d'associer ces répétitions à un facteur spécifique dans ASCA. En effet, il s'agit d'un facteur imbriqué (Marini et al., 2015). La solution proposée dans cet article consiste à utiliser des répétitions de mesures indépendantes pour réduire l'erreur de répétabilité des mesures destinées à l'analyse de variance.

3.3.3 REP-ASCA

L'analyse de variance est une opération qui peut s'interpréter dans le column-space. En effet, elle sépare les sources de variations en différents sous-espaces formés par une combinaison des observations. L'équation 3.4 correspond en fait à une série de projections orthogonales (Smilde et al., 2008). L'équation 3.4 s'écrit alors :

$$\mathbf{X} = \mathbf{M}_1\mathbf{X} + \mathbf{M}_A\mathbf{X} + \mathbf{M}_B\mathbf{X} + \mathbf{M}_{AB}\mathbf{X} + \mathbf{E} \tag{3.8}$$

Où chaque matrice M_i est une matrice de projection orthogonale du column-space ($\mathbb{R}^N$) définie par l'équation suivante :

$$\mathbf{M}_i = \mathbf{D}_i(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t \tag{3.9}$$

Où $\mathbf{D}_i$ est la matrice design appelée dummy-matrix. Cette matrice encode en matrice binaire la classe i. Comme précédemment cité, il est impossible de traiter l'erreur de répétabilité comme un facteur. En d'autres termes, il est impossible de diminuer l'effet

de cette erreur par une projection orthogonale dans le column-space. La méthode décrite dans cet article consiste à identifier les directions spectrales responsables de l'erreur de répétabilité et de les enlever de $\mathbf{X}$ par projection orthogonale dans le row-space ($\mathbb{R}^P$). Cela est réalisé en identifiant l'espace W engendré par l'erreur de répétabilité, puis en projetant $\mathbf{X}$ orthogonalement à $\mathbf{W}$ (Roger and Boulet, 2018). Supposons que des répétitions de mesures sont réalisées sur un ensemble d'échantillons représentatifs de $\mathbf{X}$. Chaque paquet de répétitions est centré puis stocké dans une matrice $\mathbf{W}$. L'information contenue dans $\mathbf{W}$ contient uniquement les effets liés à la répétition de mesure.

Puis une décomposition en valeurs singulières (de l'anglais Singular Value Decomposition ou SVD) est réalisée sur la matrice $\mathbf{W}$. Cette décomposition fournit des loadings $\mathbf{P}$ qui définissent le sous-espace $\mathbf{W}$. La projection est alors réalisée avec la formule suivante :

$$\mathbf{X}_{\perp} = \mathbf{X}(\mathbf{I} - \mathbf{P}\mathbf{P}^t) \tag{3.10}$$

L'ASCA est ensuite réalisée sur $\mathbf{X}_{\perp}$au lieu de $\mathbf{X}$. Cette méthode contient un paramètre à régler : la dimension k, c'est-à-dire le nombre de dimensions enlevées par la projection orthogonale. Si trop peu de composantes sont enlevées, alors le prétraitement de $\mathbf{X}$ aura un effet négligeable. Si trop de composantes sont enlevées (nombre de lignes de $\mathbf{P}$) et que $\mathbf{W}$ est colinéaire à l'un des facteurs, le risque est de détruire la source de variance portée par ce facteur. Le paramètre k doit donc être réglé avec précaution. Le réglage de ce paramètre peut se faire soit en examinant la forme des loadings de $\mathbf{P}$, soit en observant l'effet de la projection sur le résultat du modèle selon k.

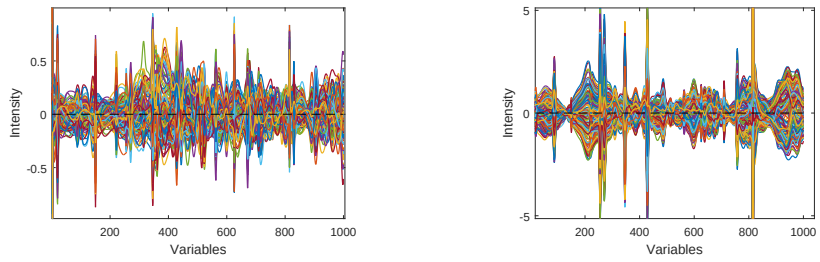
3.4 Test de REP-ASCA par simulation

L'objectif de cette partie est de valider, par la simulation, que l'approche REP-ASCA réduit bien les erreurs de répétabilités. Pour cela, des données spectrales sont générées. Ces données sont définies selon des conditions de structure typique d'une étude d'analyse de variance. C'est-à-dire que les données possèdent une structure suivant un facteur défini. Ces données sont les données de références et fournissent par la méthode ASCA des composantes principales d'intérêts.

On ajoute à ces données de références, une erreur qui vise à cacher l'expression du facteur initial. On teste, par la suite, la méthode REP-ASCA sur le jeu de données contenant des erreurs de répétabilité et on compare les résultats obtenus aux résultats de références.

3.4.1 Préparation de données simulées

Des données spectrales ont été générées (Fig. 3.3a) avec une structure définissant un facteur d'intérêt appelé A. Les niveaux de ce facteur ont été arbitrairement définis à 6 niveaux. Ce facteur exprime environ 63% de la variance totale. Le nombre de répétitions de mesure est important et est de 120 répétitions de mesures ce qui constitue 720 spectres. Un test de permutation a été réalisé pour vérifier que ce facteur soit bien significatif.



(a) (b)

FIGURE 3.3 – Spectres simulés : (a) avec le facteur A significatif (b) masquant l'effet du facteur A

Un bruit structuré a été ajouté à ces données (Fig. 3.3b). Ce bruit est défini dans un espace formé par deux composantes (Fig.3.6a). Ces composantes ont été construites en étant partiellement colinéaires aux composantes du facteur A. Des scores sont produits à partir de ces deux composantes avec une structure sous-jacente fixée par un ensemble de 6 blocs d'individus sélectionnés aléatoirement. L'ajout de ce bruit structuré permet de biaiser les tests de permutation, la variance expliquée et les loadings générés. Le facteur A est alors masqué par cette erreur et devient non significatif.

3.4.2 Résultats et discussion

Dans cette partie, nous testons REP-ASCA sur le jeu de données contenant l'erreur de répétabilité. L'objectif est alors de retrouver l'information portée par le facteur A. L'équation d'analyse de variance est alors définie par l'équation suivante :

$$\mathbf{X} = \mu + \mathbf{X}_A + \mathbf{E} \tag{3.11}$$

REP-ASCA propose d'ajouter un jeu de données indépendant visant à caractériser l'erreur de répétabilité. Dans notre cas, nous séparons aléatoirement le jeu de données simulées contenant 720 spectres. Nous utilisons ainsi 360 spectres pour décrire l'erreur de répétabilité et 360 autres spectres pour l'analyse de variance.

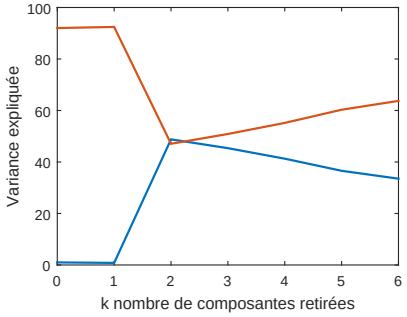


FIGURE 3.4 – Évolution de la part de variance du facteur A en fonction du nombre de composantes retirées.

La première étape est d'obtenir des composantes qui traduisent l'erreur de répétabilité. En réalisant une analyse en composante principale des paquets centrés par répétition, nous obtenons un ensemble de composantes. Le choix du nombre de composantes est motivé en étudiant un ensemble de critères. Dans notre cas, nous regardons l'évolution de la part de variance expliquée par le facteur A selon le nombre k de composantes retirées (Fig. 3.4). Sans correction (k=0), on retrouve la valeur obtenue par ASCA. Nous remarquons qu'en retirant une composante, cette valeur est inchangée. En revanche, en retirant les deux premières composantes, la valeur atteint un maximum de 47%. Nous choisissons alors de retirer les deux premières composantes.

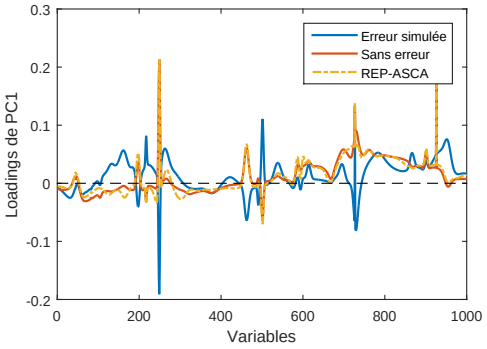


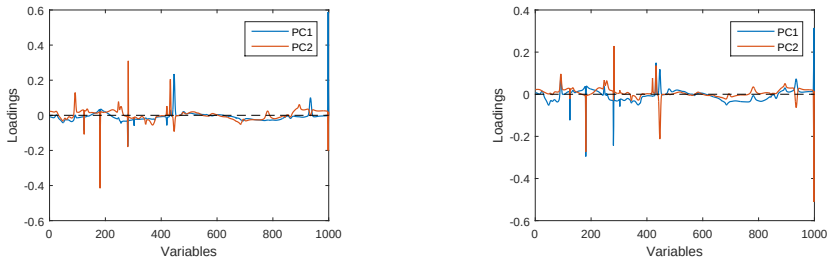
FIGURE 3.5 – Loadings obtenus sur la première composante du facteur A après (1) - courbe rouge, ASCA sur les données de références ; (2) - courbe bleu, ASCA sur les données contenant l'erreur de répétabilité, courbe en bleu ; (3) - courbe jaune, REP-ASCA sur les données contenant l'erreur de répétabilité

La figure 3.5 montre les loadings du facteur A selon trois analyse : (1) après l'analyse ASCA sur les données propres, (2) après l'analyse ASCA sur les données bruitées et (3) après REP-ASCA.

En utilisant ASCA (courbes bleu et rouge - Fig. 3.5), les composantes obtenues sont différentes lorsque les données contiennent l'erreur de répétabilité. Les pics sont décalés

et des phénomènes de pentes masquent l'information par cette composante. Cela est attendu, car l'erreur de répétabilité a été définie pour masquer l'expression du facteur A.

En revanche, en utilisant REP-ASCA, les loadings de la composante du facteur A (courbe jaune - Fig. 3.5) ressemblent à ceux des données ne contenant pas l'erreur de répétabilité après ASCA. Les pics correspondent aux pics obtenus avec ASCA sur les données de références. Il y a également beaucoup moins de décalage de pentes. La composante contient l'information non masquée par l'erreur de répétabilité. REP-ASCA permet d'obtenir la composante d'intérêt en réduisant l'erreur de répétabilité.



(a) (b)

FIGURE 3.6 – Loadings des composantes de l'erreur de répétabilité : (a) simulés (b) retrouvés avec REP-ASCA

Les deux graphiques (Fig. 3.6) montrent les composantes de l'erreur de répétabilité simulée ou obtenue par REP-ASCA. L'erreur de répétabilité identifiée par REP-ASCA est différente sur certaines régions de spectres. En revanche, REP-ASCA permet de retrouver certains pics sur les 200 premières variables et également les pentes notamment dans la région entre les variables 600 et 800.

3.4.3 Conclusion de la simulation

Un jeu de données a été généré permettant d'obtenir une composante d'intérêt associé à un facteur. Des erreurs de répétabilités ont été ajoutées à ce jeu de données de sorte à masquer le facteur étudié. Ce facteur n'est alors plus significatif et la composante porte des informations liées à l'erreur de répétabilité.

Nous avons testé REP-ASCA sur le jeu de données contenant l'erreur de répétabilité. Pour cela nous avons dû séparer le jeu de données en deux. Le premier jeu de donnée vise à caractériser l'erreur de répétabilité et le deuxième est destiné à l'analyse de variance. Nous avons étudié l'impact du nombre de composantes de l'erreur de répétabilité prise en compte pour la projection orthogonale sur la part de variance expliquée du facteur. La part de variance portée par ce facteur est maximale après avoir retiré les deux premières composantes. La composante du facteur étudié, ainsi obtenue par REP-ASCA, contient

la même information que la composante obtenue avec ASCA sur les données de référence. REP-ASCA permet bien de réduire l'erreur de répétabilité.

3.5 Conclusion du chapitre

En présence de l'erreur de répétabilité, l'interprétation des facteurs devient difficile. Les méthodes de prétraitements usuelles corrigent les lignes de base constantes. En revanche, ils ne permettent pas de réduire plus précisément l'erreur de répétabilité lorsque celle-ci s'exprime de manière plus complexe. La méthode REP-ASCA a été développée pour traiter spécifiquement les erreurs de répétabilités de mesures sur des données multivariées destinées à de l'analyse de variance. En s'appuyant sur le plan d'expérience, la description de l'erreur de répétabilité s'effectue alors à travers l'étude des répétitions de mesure. Cette description peut ainsi mettre en évidence les régions spectrales liées à l'erreur de répétabilité. Alors que cette description est réalisée dans le column-space, la réduction de cette erreur est réalisée dans le row-space en utilisant l'orthogonalisation.

Utiliser cette méthode réduit l'erreur de répétabilité en retirant les composantes principales de l'espace des répétitions. Le choix du nombre de composantes à retirer peut s'effectuer à travers un ensemble de critères tel que le critère de Durbin-Watson ou la description des loadings des composantes principales de l'espace des répétitions. De plus, ce choix peut être amélioré en étudiant précisément l'impact des projections orthogonalement à ces composantes sur les différents facteurs de l'analyse de variance.

D'un côté pratique, séparer la description de la correction de l'erreur de répétition du jeu de données permet de réduire considérablement le nombre de répétitions par modalité. En s'appuyant sur cette méthode, les protocoles d'acquisition peuvent prévoir des acquisitions visant à caractériser ces erreurs indépendamment de l'acquisition destinée à l'analyse de variance. Ce qui permet de rendre implicitement les mesures destinées à l'analyse de variance plus robustes. Il convient toutefois de conserver un nombre suffisant d'observations pour étudier la significativité des facteurs.

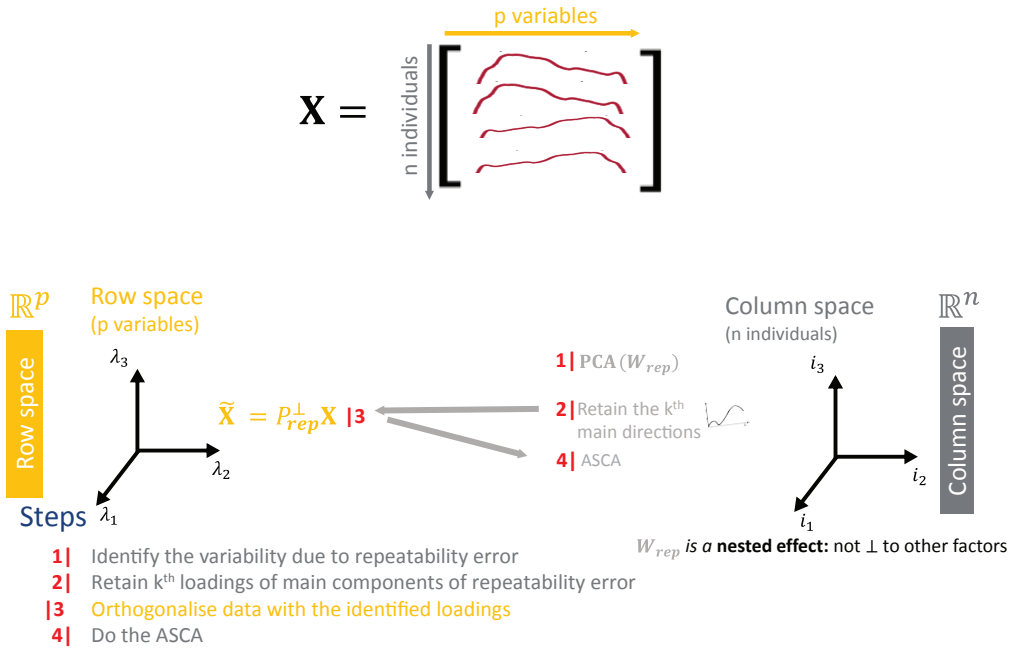


FIGURE 3.7 – Synthèse du chapitre 3 : L'illustration de la méthode REP-ASCA.

Bibliographie

Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models : a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3) :127–135, March 2009. ISSN 01695347. doi : 10.1016/j.tree.2008.10.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0169534709000196.

Marti Anderson and Cajo Ter Braak. Permutation tests for multi-factorial analysis of variance. *Journal of statistical computation and simulation*, 73(2) :85–113, 2003. URL http://www.tandfonline.com/doi/abs/10.1080/00949650215733.

Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd ; Edinburgh ; London, 1937.

JC Stanley. The Influence of Fisher's "The Design of Experiments" on Educational Research Thirty Years Later. page 8, 1962.

Gary W. Oehlert. *A first course in design and analysis of experiments*. W.H. Freeman, New York, 2000. ISBN 978-0-7167-3510-6.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 2006.

Richard G. Brereton, Jeroen Jansen, João Lopes, Federico Marini, Alexey Pomerantsev, Oxana Rodionova, Jean Michel Roger, Beata Walczak, and Romà Tauler. Chemometrics in analytical chemistry—part I : history, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry*, 409(25) :5891–5899, 2017.

A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, and M. E. Timmerman. ANOVA-simultaneous component analysis (ASCA) : a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13) : 3043–3048, July 2005. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/bti476. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti476.

J.M. Roger, F. Chauchard, and V. Bellon Maurel. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems*, 66 (2) :p. 191 – p. 204, 2003. doi : 10.1016/S0169-7439(03)00051-0. URL https://hal.archives-ouvertes.fr/hal-00464022.

Silvia De Luca, Martina De Filippis, Remo Bucci, Andrea D. Magrì, Antonio L. Magrì, and Federico Marini. Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics. *Microchemical Journal*, 129 :348–361, November 2016. ISSN 0026265X. doi : 10.1016/j.microc.2016.07.021. URL http://linkinghub.elsevier.com/retrieve/pii/S0026265X16302053.

Age K. Smilde, Huub. C.J. Hoefsloot, and Johan. A. Westerhuis. The geometry of ASCA. *Journal of Chemometrics*, 22(8) :464–471, August 2008. ISSN 08869383, 1099128X. doi : 10.1002/cem.1175. URL http://doi.wiley.com/10.1002/cem.1175.

Daniel J Vis, Johan A Westerhuis, Age K Smilde, and Jan van der Greef. Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, 8(1) :322, 2007. ISSN 14712105. doi : 10.1186/1471-2105-8-322. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-322.

Ruth G. Shaw and Thomas Mitchell-Olds. Anova for Unbalanced Data : An Overview. *Ecology*, 74(6) :1638–1645, September 1993. ISSN 00129658. doi : 10.2307/1939922. URL http://doi.wiley.com/10.2307/1939922.

Henk AL Kiers. *SCA : A Program for Simultaneous Components Analysis of Variables Measured in Two Or More Populations : user's Manual*. iec ProGamma, 1990.

Katrijn Van Deun, Age K. Smilde, Mariët J. van der Werf, Henk AL Kiers, and Iven Van Mechelen. A structured overview of simultaneous component based data integration. *Bmc Bioinformatics*, 10(1) :246, 2009.

Phil Williams and Karl Norris. *Near-infrared technology in the agricultural and food industries.* American Association of Cereal Chemists, Inc., 1987.

Federico Marini, Dalene de Beer, Elizabeth Joubert, and Beata Walczak. Analysis of variance of designed chromatographic data sets : The analysis of variance-target projection approach. *Journal of Chromatography A*, 1405 :94–102, July 2015. ISSN 00219673. doi : 10.1016/j.chroma.2015.05.060. URL http://linkinghub.elsevier.com/retrieve/pii/S0021967315007839.

Jean-Michel Roger and Jean-Claude Boulet. A review of orthogonal projections for calibration. *Journal of Chemometrics*, page e3045, June 2018. ISSN 08869383. doi : 10.1002/cem.3045. URL http://doi.wiley.com/10.1002/cem.3045.

Potentiel d'une forte résolution spectrale pour la sélection variétale

Chapitre 4

Sommaire

4.1 Introduction **78**

4.2 Matériels et méthodes **78**

4.2.1 Dispositif expérimental 78

4.2.2 Données météorologiques et agronomiques collectées 79

4.2.3 Mesures spectrales 79

4.2.4 Protocole de mesures spectrales 81

4.3 Résultats et discussion **81**

4.3.1 Description des variables agronomiques 82

4.3.2 Caractérisation des génotypes sur la base des données agronomiques 83

4.3.3 Caractérisation des conditions climatiques 84

4.3.4 Application de REP-ASCA 86

Terme traitement 89

Terme génotype 91

Terme interaction 93

4.4 Recherche de traits phénotypiques **96**

4.4.1 Terme traitement 96

4.4.2 Terme génotype 97

4.4.3 Terme interaction 98

4.5 Conclusion du chapitre **100**

4.1 Introduction

L'information spectrale permet d'accéder aux caractéristiques physico-chimiques de la végétation. Une forte résolution spectrale est idéale pour le phénotypage des plantes (Vigneau, 2010; Cabrera-Bosquet et al., 2012; Araus and Cairns, 2014). L'utilisation de données spectrales permet le suivi de culture (Ecarnot et al., 2013; Mahajan et al., 2016; Vigneau et al., 2011), la détection de maladies (Mahlein et al., 2012), ou la détection précoce de stress biotiques ou abiotiques (Behmann et al., 2014; Römer et al., 2012). Des études ont montré qu'il était possible d'utiliser cette information en sélection variétale notamment pour la prédiction du rendement (Montesinos-López et al., 2017; Aguate et al., 2017) ou pour accéder à d'autres traits d'intérêt agronomique (surface foliaire, biomasse) par le biais de combinaison de quelques bandes spectrales (Aparicio et al., 2000; Babar et al., 2006).

Dans ce chapitre nous allons montrer que des mesures spectrales permettent de discriminer finement des génotypes dans le cadre de la sélection variétale. Nous montrerons comment utiliser des données spectrales comme traits phénotypiques pour décrire les réponses des variétés face à un stress hydrique. Le spectre de réflectance entier peut être utilisé comme proxi contenant un ensemble d'informations biochimiques. Nous nous appuierons sur la méthode REP-ASCA pour tirer parti du plan d'expérience associé aux données spectrales.

Premièrement nous détaillerons le dispositif expérimental et le protocole de mesure utilisé. Puis, nous appliquerons REP-ASCA à une des dates des mesures spectrales. Nous décrirons les résultats et détaillerons l'utilité d'une telle méthode pour exploiter des données spectrales conjointement avec un plan d'expérience.

4.2 Matériels et méthodes

4.2.1 Dispositif expérimental

Deux campagnes expérimentales ont été menées. La première a eu lieu en 2017 sur le site d'essai de Limagrain situé à Aubiat (63). La deuxième a eu lieu en 2018 sur la commune de Nérac (47). Le même plan expérimental a été utilisé pour les deux campagnes. Ce plan était composé d'un ensemble de 10 génotypes commerciaux identifiés par des lettres de A à J. Ces génotypes ont été choisis pour avoir des tolérances contrastées au stress hydrique. Sur la base d'un grand nombre d'essai, les génotypes A et C sont classés

comme très sensibles et les génotypes E, F et B sont tolérants.

Le plan expérimental était organisé en bloc [1] de manière aléatoire. Le plan était complet [2] et équilibré [3]. Chaque génotype était présent en 2 répétitions sous 2 conduites d'irrigation. Cela représentait un dispositif de 40 micro-parcelles. Les deux traitements d'irrigation étaient les suivants : en condition d'irrigation optimale et un autre sans irrigation donc en apport d'eaux pluviales. Chaque parcelle a été semée en 4 rangées avec une densité de 8.5 plantes/m^2 en 2018 pour les deux conditions d'irrigation et des densités de 8.5 plantes/m^2 en condition non irriguée et de 9.6 plantes/m^2 en condition irriguée en 2017. Chaque micro-parcelle a été identifiée par sa position XY dans la parcelle.

4.2.2 Données météorologiques et agronomiques collectées

Deux stations météorologiques ont été installées sur le terrain pour déterminer localement la température, la lumière et l'hygrométrie. Pour corroborer l'étude, des mesures traditionnelles ont été effectuées telles que des mesures de rendement, d'humidité du grain et des notations (qualité de parcelles et date de floraison) ou les vols de drones (NDVI). De plus, des tensiomètres (sondes Watermark) ont été installés sur les deux niveaux de traitement. Des mesures de tensions dans le sol ont alors été réalisées pour décrire les réserves hydriques des deux conditions d'irrigation. Des mesures de déficit de pression de vapeur ou Vapour Deficit Pression (VPD) ont été également obtenues à partir de l'écart entre la pression d'eau saturante (hygrométrie ou humidité relative) et la pression de vapeur d'eau caractérisant l'air (hygrométrie ou température). Ce déficit permet de caractériser la demande évaporative.

4.2.3 Mesures spectrales

Un dispositif a été conçu pour pouvoir réaliser des mesures spectrales de la végétation du maïs. L'ensemble du dispositif contenait les éléments suivants :

— un spectromètre Zeiss MMS1 d'une gamme spectrale allant de 310 nm à 1100 nm et d'une résolution spectrale de 3 nm

— une sphère d'intégration, ou sphère intégrante servant à mesurer le spectre du rayonnement incident sur 2π sr.

1. Un bloc est aussi appelé unité expérimentale qui correspond au plus petit élément du plan d'expérience.

2. Un plan expérimental est dit complet lorsque tous les blocs sont représentés dans l'ensemble des facteurs ou traitements appliqués.

3. Un plan est dit équilibré lorsque le nombre d'unité expérimentale est le même pour tous les niveaux des facteurs.

— un ordinateur pour piloter les acquisitions

— une batterie Li-Po avec une autonomie d'environ 24 h

— un support (Fig. 4.2) pour effectuer des mesures au-dessus de la canopée du maïs pour tous les stades de développement.

Le schéma 4.1 représente le dispositif destiné à mesurer les spectres. La mesure avec la sphère d'intégration orientée vers le zénith[4] a été réalisée pour mesurer le spectre du rayonnement incident. Ce spectre est utilisé pour calculer les spectres en réflectance et donc de normaliser les spectres selon les changements de luminosité. Une fibre a été dirigée vers le bas pour pouvoir mesurer la végétation. Un switch manuel permettait de commuter entre les mesures effectuées via la sphère d'intégration et les mesures issues de la fibre.

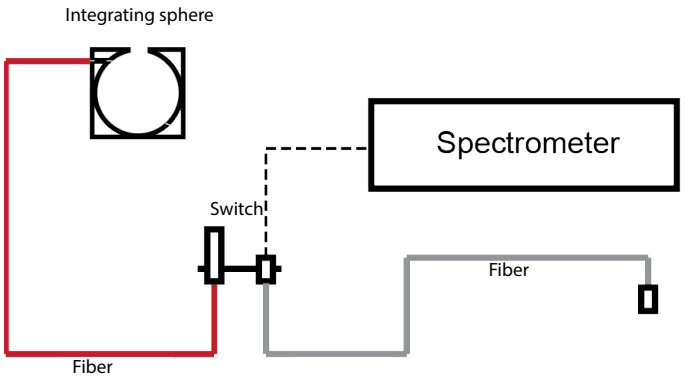


FIGURE 4.1 – Dispositif de mesure des spectres.

FIGURE 4.2 – Support de mesure.

4. Le zénith est le point orienté vers le ciel.

4.2.4 Protocole de mesures spectrales

Une série de 12 mesures de réflectance a été réalisée pour chaque micro-parcelle. Les mesures ont été effectuées sur une des deux rangées centrales de chaque micro-parcelle. Les mesures sont illustrées figure 4.3. Le parcours est défini selon l'orientation du champ et de l'heure de l'acquisition afin de minimiser les ombres portées dues au dispositif de mesure. Une mesure de référence avec la sphère d'intégration a été acquise systématiquement pour chaque mesure sur la végétation. Cela représentait donc un ensemble de 24 spectres par micro-parcelle.

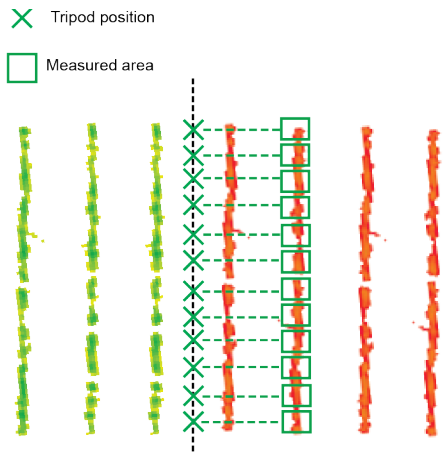


FIGURE 4.3 – Points de mesures établis sur une micro-parcelle semée en 4 rangs (rouge).

Pour utiliser REP-ASCA, 8 mesures par micro-parcelle sont choisies aléatoirement pour l'analyse de variance. Les points de mesures restants constituent une série supplémentaire qui permet de caractériser l'erreur de répétabilité.

Avant de commencer l'acquisition sur la parcelle, le temps d'intégration du spectromètre a été fixé. Ceci a été défini manuellement en analysant préalablement les valeurs maximales obtenues pour les spectres mesurés. Pour bénéficier de toute la plage du capteur, la valeur maximale du spectre mesuré a dû être comprise entre 60% et 80% de la valeur maximale autorisée par le spectromètre.

Des mesures ont été réalisées sur plusieurs dates pour chaque campagne : 7 dates pour la campagne de 2017 et 6 dates pour la campagne de 2018. Ces dates ont été choisies pour couvrir l'ensemble des stades de développement du maïs. Dans ce chapitre nous étudions uniquement les données de la campagne de 2018.

4.3 Résultats et discussion

Il est difficile d'établir un critère unique pour caractériser la résistance d'un génotype face à un stress (Farshadfar and Sutka, 2002). Dans notre étude, les pourcentages de

pertes en rendement entre les deux conditions d'irrigation sont utilisés pour caractériser les réponses des génotypes vis-à-vis du stress hydrique. Pour un génotype, si le pourcentage de perte est proche de 0%, le génotype montre peu de sensibilité face à la différence de traitement. Au contraire, un pourcentage de perte élevé montrera une sensibilité face au changement de condition d'irrigation. La valeur de rendement de la condition irriguée est aussi importante. En effet, elle montre le rendement potentiel des génotypes sous conditions optimales d'irrigation.

4.3.1 Description des variables agronomiques

TABLE 4.1 – Moyenne, écart type et héritabilité des variables rendement et humidité obtenues à la récolte pour les deux conditions d'irrigation.

Variable	Irrigué			Non irrigué		
	Moyenne	écart-type	H2	Moyenne	écart-type	H2
Rendement (q/ha)	104.42	12.48	0.58	85.46	9.12	0.24
Humidité (%)	14.54	1.53	0.98	13.38	1.20	0.82

La table 4.1 montre par traitement, les valeurs moyennes, les écarts-types et les héritabilités des valeurs de rendements et d'humidité du grain obtenues à la récolte. L'héritabilité H2 est définie au sens large (Kruijer et al., 2015) comme la variabilité génotypique sur la variabilité phénotypique. Elle se calcule selon l'équation suivante :

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{4.1}$$

Où σ_g^2 est définie par l'équation : $\sigma_g^2 = \frac{MS(G)-MS(E)}{r}$, où r représente le nombre de répétitions, $MS(G)$ et $MS(E)$ représentent les moyennes de sommes des carrées issues d'une analyse de variance pour le terme du génotype et l'erreur résiduelle. σ_e^2 est défini par l'équation : $\sigma_e^2 = MS(E)$. Si la valeur d'héritabilité d'une variable mesurée est proche de 1, cette variable mesurée dépendra fortement du génotype. Elle constituera donc un bon descripteur du génotype. En revanche, si l'héritabilité de la variable mesurée est proche de 0, celle-ci dépendra uniquement de l'environnement et ne sera donc pas robuste. Elle ne pourra pas être utilisée pour décrire les génotypes.

En condition irriguée, le rendement moyen est de 104.42 q/ha avec un écart-type de 12.48 q/ha et une héritabilité de 0.58. En condition non irriguée, le rendement moyen est de 85.46 q/ha. L'héritabilité, de 0.24 est beaucoup plus faible dans la condition non irriguée. Pour l'humidité du grain, la moyenne est de 13.96%, l'écart type de 1.48% et l'héritabilité de 0.84. Pour la condition irriguée, la moyenne est de 14.54% avec un

écart-type de 1.53%. L'héritabilité est de 0.98. Dans la condition non irriguée, la moyenne d'humidité est de 13.38% avec un écart-type de 1.2 et une héritabilité de 0.82.

Obtenir des valeurs inférieures dans la condition non irriguée de la condition irriguée était attendu. Le rendement en condition non-irriguée est réduit de près de 18 % par rapport à la condition irriguée. Concernant l'héritabilité, les valeurs obtenues pour l'humidité étaient attendues. En effet, c'est un trait fortement héritable quel que soit le traitement. En revanche, le rendement est moins héritable. L'héritabilité en condition irriguée est supérieure à celle de la condition non irriguée. Ceci s'explique par la diminution rapide de la variance génétique par rapport à la variance environnementale face à un stress qui augmente (Bolanos and Edmeades, 1996). En effet, la variance environnementale augmente avec l'apparition du stress hydrique qui entraîne des hétérogénéités parcellaires.

Table 4.2 – ANOVA sur la variable rendement.

Facteur	d.f.	Sum Sq.	Mean Sq.	F	p-value
Traitement	1	3594.36	3594.36	55.85	0
Genotype	9	2138.6	237.62	3.69	0.007
Interaction	9	1113.51	123.72	1.92	0.10
Residual	20	1287.14	64.36		
Total	39	8133.61	85.5894		

La table 4.2 montre les résultats d'une analyse de variance sur la variable rendement. Les p-values pour les facteurs traitement et génotype sont inférieures à 0.01. La p-value du terme d'interaction est de 0.10. Les facteurs génotype et traitements sont significatifs pour ces deux variables mesurées.

Les rendements en grains ont donc été significativement réduits en traitement non irrigué par rapport aux conditions bien arrosées. Dans cette étude, il y a peu de génotypes et de parcelles, la taille de l'essai est donc très faible. Les résultats obtenus pour les p-values des facteurs génotype et traitement sont attendus. Les génotypes ont été choisis pour avoir des comportements très différenciés vis-à-vis d'un déficit hydrique.

4.3.2 Caractérisation des génotypes sur la base des données agronomiques

Pour chaque génotype, le rendement moyen pour les deux traitements et le pourcentage de pertes sont fournis dans la table 4.3. Pour la variable rendement dans la condition irriguée, certains génotypes ont des valeurs très élevées comme le génotype G et C avec des valeurs moyennes de 115.3 q/ha et 118.5 q/ha. D'autres ont des valeurs très basses comme les génotypes E et J avec des valeurs de 78.39 q/ha et 99.68 q/ha.

TABLE 4.3 – Rendement moyen par traitement et pourcentage de pertes pour chaque génotype.

Genotype	Rendement (q/ha)		
	Irr	Non-irr	Loss (%)
A	103.22	83.17	19.42
B	105.85	97.01	8.34
C	115.39	88.92	22.94
D	100.82	93.30	7.45
E	78.39	77.13	1.61
F	103.63	92.49	10.74
G	118.55	86.65	26.91
H	104.16	76.29	26.75
I	114.51	81.64	28.70
J	99.68	77.99	21.75

En condition non irriguée, les génotypes possédant de fortes valeurs en rendement sont les génotypes B et D avec respectivement 97.02 q/ha et 93.31 q/ha. Les génotypes dont les rendements sont les plus faibles sont les génotypes E et J avec des valeurs respectives de 77.13 q/ha et 77.99 q/ha.

Les génotypes qui ont un faible pourcentage de pertes en non irrigué sont les génotypes B, D et E avec respectivement une perte de 8.34, 7.45% et 1.61%. Les génotypes très sensibles sont les génotypes G, H et I avec respectivement des pertes de 26.90% , 26.75% et 28.70%.

Sur la base du rendement, la tolérance des génotypes face au stress hydrique peut être décrite ici par le pourcentage de pertes de rendement tandis que la valeur de rendement en condition irriguée traduit le potentiel de rendement du génotype en conditions optimales. À partir de ces variables, il est possible d'établir deux classes de génotypes. La première classe de génotype regroupe les génotypes tolérants dont les génotypes B, D et E tout en dissociant le génotype E qui présente des rendements très faibles et B et D des rendements plus élevés. La deuxième classe est composée de génotypes sensibles comme les génotypes G, H et I. Pour dissocier ces génotypes au sein de cette classe, G et I possèdent des rendements en condition irriguée très élevés tandis que le génotype H possède de faibles rendements.

4.3.3 Caractérisation des conditions climatiques

La date choisie doit correspondre à une des dates pour laquelle il y a le plus de différence entre les deux traitements avec si possible une apparition de stress hydrique pour la partie non irriguée. Le stress hydrique est une combinaison d'un manque d'eau

dans le sol, d'une demande évaporative élevée et de températures élevées. Il convient alors de choisir une date où la température maximale, la demande évaporative et le manque d'eau dans le sol sont élevés. Pour cela, des informations climatiques sont disponibles pour chaque date d'acquisition spectrale. Les informations retenues pour cette étude sont : les températures maximales et la tensiométrie pour chacun des traitements et du déficit de pression de vapeur (VPD). Ces informations sont fournies dans les tables 4.4 et 4.5. Les températures maximales des 3 jours qui précédent la date d'acquisition sont également fournies dans la table 4.5.

Les valeurs de VPD sont relativement hautes sur l'ensemble des dates. Le maximum est atteint à la première date avec une valeur de 1.60 kPa. Pour les dates 2,3,4,5 les valeurs sont situées entre 1.19 kPa et 1.30 kPa. Pour la dernière date, la VPD est plus faible avec une valeur de 0.83 kPA. Les valeurs hautes signifient une demande évaporative élevée. L'air est sec et accroît la transpiration des plantes. Une valeur basse signifie que la demande évaporative de l'air est faible et que la transpiration est moins importante. Hormis la dernière date, l'ensemble des dates d'acquisitions présentent une demande évaporative élevée.

Les températures sont données table 4.5. Les températures maximales varient selon les dates d'acquisitions. La température maximale est très forte dès la première date avec une valeur de 31.6°C. La deuxième valeur la plus élevée est atteinte à la date 3 avec 30.8 °C. Dans les 3 jours qui précédent les dates d'acquisition, la température maximale de la première date est plus faible de 3°C. En revanche, pour les autres dates, les températures maximales sont du même ordre de grandeur qu'aux dates d'acquisitions.

Ces températures sont celles attendues dans le sud-ouest de la France (Parey et al., 2007). Les valeurs obtenues sur les trois jours qui précèdent confirment que les températures des jours des acquisitions ne sont pas anormales sur la période. Les jours correspondants aux mesures spectrales ne sont pas atypiques des jours qui ont précédé.

Les valeurs tensiométriques changent selon la date d'acquisition spectrale. Pour la partie non irriguée, les valeurs sont relativement hautes et supérieures à 220 cbar pour les dates 3,5 et 6. Les valeurs pour les dates 2 et 4 sont plus faibles et aux alentours de 145 cbar. La valeur tensiométrique pour la partie irriguée augmente progressivement tout au long des dates de mesure, passant de 78,5 cbar à 168 cbar. Le différentiel entre les deux traitements est maximal à la troisième date avec une différence de 153 cbar et est plus faible pour les dates suivantes. Le différentiel est minimal à la date 4. Les valeurs de tensions obtenues pour la partie non irriguée sont celles attendues pour des essais en condition de stress hydrique. Des pluies ont eu lieu entre la date 3 et la date 4 ce qui explique que le différentiel de tensions dans le sol entre les traitements diminue fortement à la date 4. La troisième date correspond à la date où le différentiel est le plus

important.

TABLE 4.4 – Déficit de pression de vapeur (VPD) et tensiométrie (cbar) des dates d'acquisitions de l'année 2018.

N°	Date	VPD(kPA)	P(non-irr)(cbar)	P(irr)(cbar)	Difference
1	26-Jun-18	1.60	x	x	x
2	10-Jul-18	1.19	145.08	x	<145.08
3	30-Jul-18	1.25	231.5	78.5	153
4	20-Aug-18	1.30	145.17	86.67	58.5
5	27-Aug-18	1.23	229.5	121	108.5
6	10-Sep-18	0.83	239	168	71

TABLE 4.5 – Températures (°C) maximales de la journée et pour les 3 jours précédents des dates d'acquisitions de l'année 2018.

N°	Date	Tmax(°C)	Tmax(-3j)
1	26-Jun-18	31.6	29.1
2	10-Jul-18	28.8	31.1
3	30-Jul-18	30.8	34.2
4	20-Aug-18	29.5	29.9
5	27-Aug-18	29.8	30.6
6	10-Sep-18	26.9	29.4

La date 3 correspond à une valeur tensiométrique élevée pour la condition non irriguée, combinée à une demande évaporative moyenne mais à des températures élevées. Pour la suite de l'étude, cette date est donc choisie afin d'obtenir des différences notables entre les deux conditions d'irrigation. De plus, cette date correspond au stade du remplissage du grain qui est un des stades les plus sensibles. En effet, à cette date la sécheresse a un impact sur le remplissage du grain et donc sur la valeur de poids de mille grains (PMG) (Chapuis et al., 2012; Aslam et al., 2015).

4.3.4 Application de REP-ASCA

Selon les résultats précédents, l'analyse de variance REP-ASCA est effectuée sur la date 3 où on est susceptible de voir des différences entre les traitements.

La figure 4.4 montre l'ensemble des spectres acquis à la date 3. Nous remarquons que les valeurs de réflectance sont plus élevées dans le domaine spectrale du proche infrarouge que dans le visible. Ces spectres de réflectance correspondent donc bien à des spectres de végétation. On constate qu'ils sont affectés par un effet de variation systématique. Cet effet est composé d'une translation verticale, visible vers 400 nm et

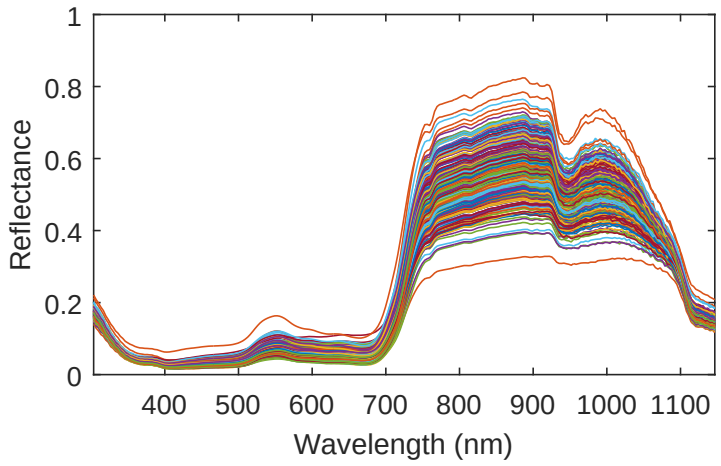


FIGURE 4.4 – Spectres de réflectance des mesures spectrales effectuées à la date 3 de l'année 2018.

d'un effet multiplicatif, puisque les variations verticales sont plus importantes entre 800 et 1000 nm, là où la valeur du spectre est la plus grande.

Les effets multiplicatifs sont classiquement observés sur les mesures en réflectance de la végétation. Ils sont dus principalement aux variations des angles formés entre la normale de la surface des feuilles et l'axe de mesure (Vigneau, 2010). Les effets additifs sont plus mineurs. Ils sont généralement attribués à la présence de points de réflexion spéculaire (Vigneau, 2010) Les effets additifs et multiplicatifs sont généralement traités par l'application de méthodes de normalisation, telles que SNV (Barnes et al., 1989). Cependant, ces prétraitements peuvent avoir des effets délétères sur les spectres (Bi et al., 2014, 2016). C'est pourquoi aucun prétraitement n'a été appliqué ici.

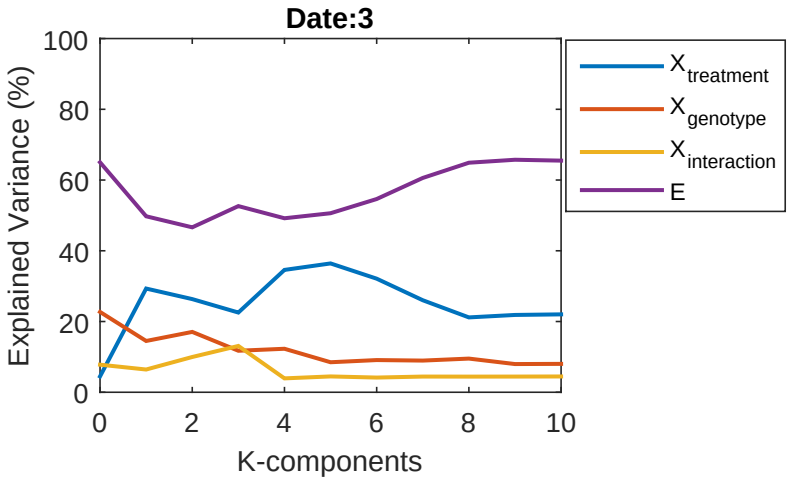


FIGURE 4.5 – Impact des projections sur les variances expliquées des facteurs de l'analyse de variance.

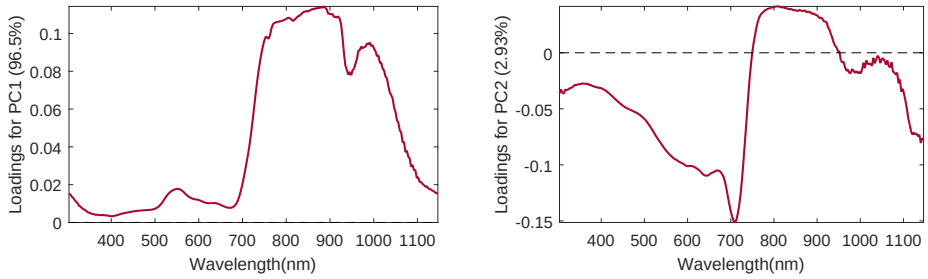
En utilisant REP-ASCA, le protocole de mesure prévoit de mesurer un jeu de données permettant de caractériser l'erreur de répétabilité. Ce jeu de données permet d'estimer le sous-espace porteur des perturbations spectrales dues aux erreurs de répétabilité puis de retirer ce sous-espace par projection orthogonale sur les observations destinées cette fois-ci à l'analyse de variance. Pour cela, la dimension k de ce sous-espace doit être déterminée.

La figure 4.5 montre l'évolution des pourcentages de variance expliquée de l'ensemble des facteurs étudiés en fonction de k. Sur cette figure, les projections orthogonales font varier les pourcentages de variance expliquée des facteurs. Ces variations ne se font pas dans les mêmes proportions. Pour le facteur traitement, sans apporter de correction (lorsque k=0), la variance expliquée par ce facteur est de 4.45 % de la variance totale. Après la projection orthogonale à la première composante (k=1), la variance expliquée par ce facteur passe à 29.32 % ce qui représente la plus grande augmentation en effectuant une projection. La variance varie par la suite selon les projections entre 22 % et 36 % pour k>1. Pour le facteur génotype, la variance expliquée par ce terme diminue de manière à peu près constante selon k. Pour le terme d'interaction, la variance expliquée augmente légèrement jusqu'à k=3 pour atteindre 13.09 %. Mais à partir de 4 projections, la variance expliquée chute fortement pour atteindre entre 3 et 4 % de la variance totale. Sans correction, la part de variance expliquée du terme résiduel est très élevée et atteint 65.01 % de la variance totale. Elle est toujours supérieure aux parts de variance expliquée des autres facteurs. Cependant, elle diminue avec les projections orthogonales jusqu'à k=2 pour atteindre le minimum à 46.63 % de la variance totale. Par la suite, cette part de variance augmente progressivement.

La première projection permet de réduire fortement la part de variance résiduelle et la part de variance du facteur traitement augmente fortement. Cette première composante masque l'expression du terme traitement. Pour l'ensemble des autres facteurs (génotype et interaction), les projections orthogonales aux deux premières composantes ont beaucoup moins de conséquences sur les pourcentages de variances expliquées.

Puisqu'à k=4, le terme d'interaction ne s'exprime plus, le nombre k doit être inférieur à 4. À k=3, la part de variance résiduelle est très élevée. Le choix s'est donc porté à k=2. En effet, cela représente le moment où la part de variance résiduelle est la plus petite.

Une SVD est effectuée sur la matrice $\mathbf{W}$ contenant l'erreur de répétabilité (Section 3.3.3). Les loadings de la première composante sont visibles figure 4.6a. Ils représentent 96.5 % de variance de l'erreur de répétabilité. Ces loadings ressemblent au spectre moyen des spectres obtenus (Fig. 4.4). De plus, tous les loadings sont de même signe. Cela correspond à un effet systématique, additif ou multiplicatif, dont les variations sont



(a) (b)

FIGURE 4.6 – Description de l'erreur de répétabilité : (a) loadings de la première composante de $\mathbf{W}$, (b) loadings de la deuxième composante de $\mathbf{W}$.

alignées avec le spectre moyen. On constate donc qu'une grande part de l'erreur de répétabilité est portée par les effets systématiques décrits précédemment (Fig. 4.4).

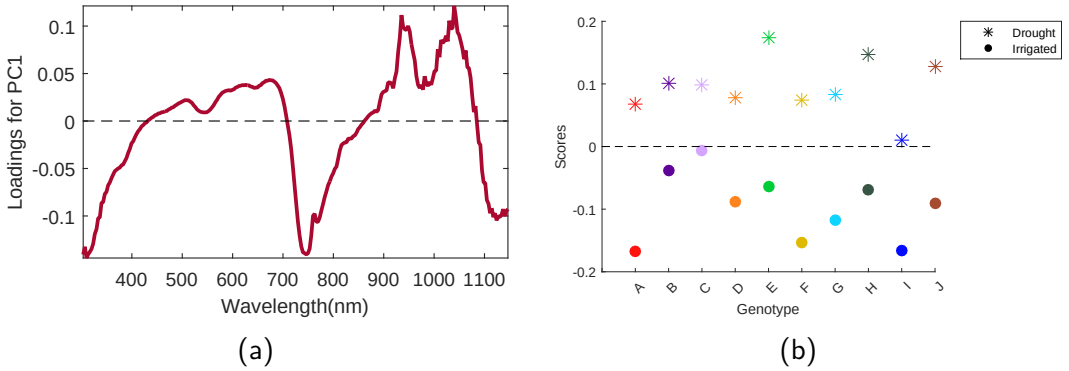
Les loadings de la deuxième composante sont visibles en figure 4.6b. Ces loadings montrent une opposition entre la région visible (300 - 700 nm) et la partie proche infrarouge (700 – 1100 nm). De plus, ils présentent un pic prononcé, à 709 nm. L'erreur de répétabilité au sein d'une micro-parcelle s'exprime ici comme une différence de pente entre la partie visible sensible aux pigments et la partie proche infrarouge sensible à la structure des cellules. Le pic montre une variation de la position du red-edge et plus particulièrement sur le début du red-edge (Boochs et al., 1988).

Il semble donc que l'erreur de répétabilité affecte également la forme du spectre de réflectance. Il est connu (Gausman and Allen, 1973; Gausman et al., 1974) que la balance entre les parties visible et proche infrarouge, ainsi que la position du red-edge peuvent varier d'une feuille à l'autre de la même plante et par conséquent d'une plante à l'autre. Il n'est donc pas surprenant de retrouver ces déformations dans l'erreur de répétabilité.

L'erreur de répétabilité du jeu de données ayant été réduite, l'ASCA est réalisée sur le jeu de données corrigé. L'ASCA fournit ainsi des loadings et des scores pour chaque terme. Le terme traitement est un facteur à deux niveaux. Nous obtenons donc une seule composante. Pour les termes génotypes et les termes d'interaction, nous détaillerons les résultats obtenus pour les deux premières composantes.

Terme traitement

Les loadings de la composante principale sont présentés sur la figure 4.7. On observe dans les parties visible et proche infrarouge deux massifs de même signe et séparés par un creux centré sur 739 nm. Une observation plus détaillée montre deux pics positifs situés à 945 nm et 1040 nm. Une pente est également visible au début de la composante jusqu'à 420 nm. Le creux situé à 739 nm correspond à la région du red-edge et plus précisément



(a) (b)

FIGURE 4.7 – Description du terme traitement : (a) Loadings de la composante principale, (b) Scores obtenus par génotype sur la composante principale.

à son point d'inflexion (maximum de la première dérivée). Les loadings étant négatifs à 739 nm, plus la courbure sera faible plus le score résultant sera positif. En d'autres termes, plus le spectre présentera un red-edge raide et plus le score obtenu sera négatif. La pente entre 300 et 420 nm exprime un changement dans la région des ultraviolets (UV). Les UV sont aussi connus pour être un indicateur de stress (Crafts-Brandner and Salvucci, 2002; Lichtenthaler and Miehé, 1997).

D'un autre côté, les deux pics situés à 935 nm et 1040 nm entourent le creux situé à 970 nm correspondant à l'absorption de l'eau (Osborne et al., 1993; Ustin et al., 2012). La présence de ces deux pics montre un effet d'élargissement du creux situé à 970 nm. Plus le score correspondant est élevé, plus les régions spectrales qui entourent ce creux sont élevées. L'absorption liée à l'eau sera donc moins élargie. En résumé, les loadings associés au facteur traitement permettent de produire des scores négatifs à partir des spectres au red-edge marqué et présentant un élargissement plus prononcé au niveau du creux lié à l'absorption de l'eau.

Les scores obtenus pour chaque génotype, pour chaque traitement et après projection des observations sur cette composante sont présentés figure 4.7b. Les scores obtenus en condition irriguée (symbole rond) sont principalement négatifs et les scores en condition non irrigués (symbole étoile) sont majoritairement positifs. On constate que le génotype E en condition non irriguée est très éloigné des autres génotypes. On remarque également que le génotype I possède un score (en condition irrigué) proche de zéro. Ces scores fournissent des informations sur les caractères liés au génotype en réponse au comportement moyen par traitement. Les génotypes dans des conditions irriguées ont alors un point d'inflexion de la pente du red-edge plus marqué et une absorption plus importante au niveau de l'eau. Le changement en teneur en eau des feuilles peut exprimer des changements dans le métabolisme des plantes pour mettre en place des mécanismes face à la sécheresse (Kavar et al., 2008). Certains génotypes, comme le génotype I en

conditions non irriguées ou le C en conditions bien irriguées ne respectent pas cette règle. En effet, les scores sont proches de zéro. Cela signifie que ces génotypes sont éloignés des comportements moyens pour un traitement donné. On peut émettre l'hypothèse que dans le cas non irrigué, les génotypes I et C ne mettent pas en place une réponse face au stress subi.

De la littérature, on sait que sous un stress léger, la fermeture des stomates joue un rôle protecteur en réduisant la perte d'eau et en limitant les échanges gazeux (Aslam et al., 2015; Chaves et al., 2003). Cette fermeture stomatique entraîne directement une pénalité de rendement, car elle agit directement sur le processus photosynthétique même lors d'un stress léger de sécheresse. Sous stress sévère la fermeture a des conséquences importantes sur le système photosynthétique (Chaves et al., 2009). Ce terme nous permet de décrire ce qui différencie les deux traitements via des mécanismes communs pour l'ensemble des génotypes. Les scores par génotype et par traitement nous permettent alors de situer les génotypes par rapport au comportement commun. Certains génotypes comme le génotype C et I semblent atypiques.

Terme génotype

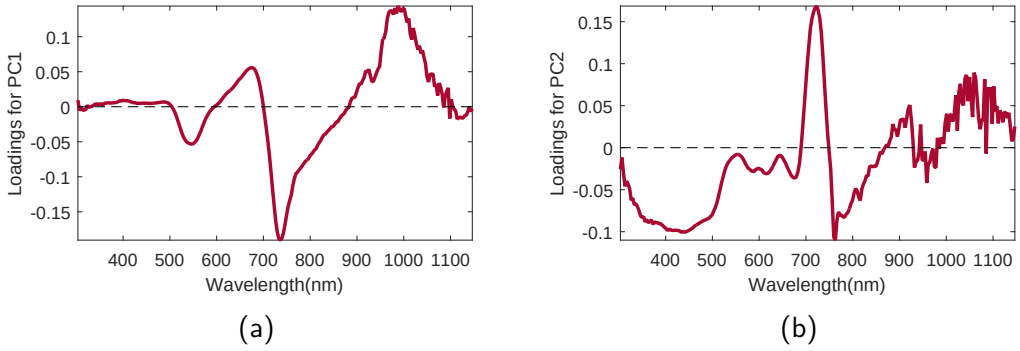


FIGURE 4.8 – Description du terme génotype : (a) Loadings de la première composante principale, (b) Loadings de la deuxième composante principale.

Les loadings des deux premières composantes de ce terme sont illustrés en figure 4.8. Sur la première composante (Fig. 4.8a), les loadings sur la plage spectrale 300-500 nm sont nuls. Deux creux (pics négatifs) sont situés à 545 et à 736 nm. Des pics sont situés à 672 nm et à 980 nm. Une cassure à partir de 769 nm intervient dans la pente croissante visible de 736 à 980 nm. Les creux situés à 545 nm et à 736 nm correspondent respectivement à la teneur en anthocyanes et au point d'inflexion du red-edge. Les pics de 672 et 980 nm correspondent respectivement à l'absorption de la chlorophylle et de l'eau. La première composante oppose alors les observations selon les teneurs en chlorophylle et en eau à la teneur en anthocyanes et de la pente du le red-edge. Un score

négatif traduira une forte teneur en eau et en chlorophylle et décrira donc un bon état de la plante et une bonne activité photosynthétique. À l'inverse un score positif traduira une faible capacité photosynthétique avec l'apparition de stress environnementaux par la présence d'anthocyanes ou une pente de red-edge moins importante. Cette composante décrit globalement la capacité photosynthétique.

La forme des loadings de la deuxième composante (Fig. 4.8b) montre une partie négative constante entre 350 nm et 500 nm, un pic à 719 nm et une pente constante de 760 nm à 1080 nm. La partie négative de 350 nm à 500 nm correspond aux pigments et plus particulièrement aux caroténoïdes et chlorophylles. Le pic situé à 719 nm correspond à la première partie du red-edge. Cette composante apporte des informations complémentaires à la première composante et concernant plus spécifiquement des teneurs en caroténoïdes. Ces pigments peuvent protéger le système photosynthétique de l'excès de lumière (Demmig-Adams, 1990; Gitelson et al., 2002). Des scores positifs traduisent ici une réflectance plus faible à 500 nm et donc une teneur plus élevée en caroténoïdes. Les génotypes ayant une teneur élevée en caroténoïdes peuvent être retenus pour leurs capacités à être tolérant face à l'excès de lumière ou d'ombre (Pessarakli, 1999).

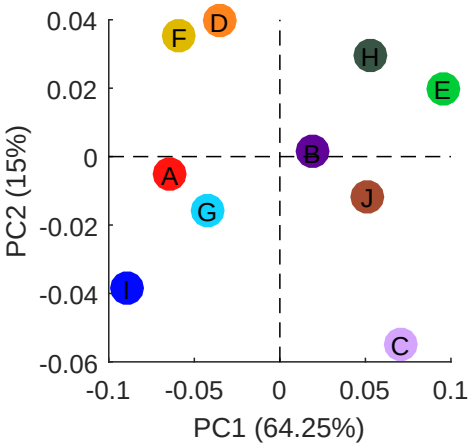
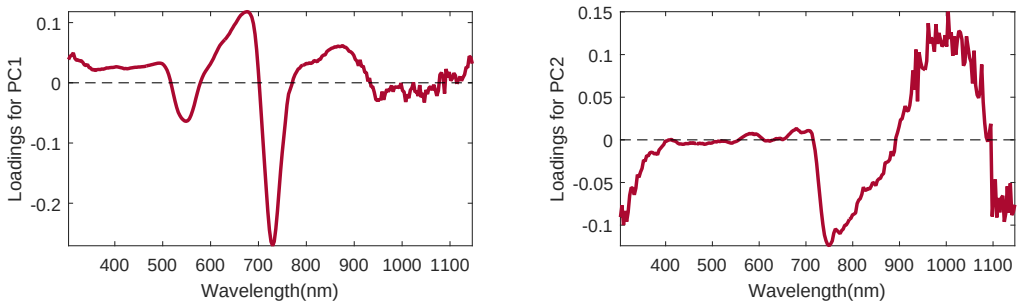


FIGURE 4.9 – Scores des génotypes sur les deux premières composantes (PC1 et PC2) du terme génotype.

Le scores-plot obtenu sur les composantes principales 1 et 2 (PC1 et PC2) est illustré figure 4.9. Certains génotypes se regroupent : les génotypes F et D ; les génotypes A , I et G ; les génotypes E et H. Ce regroupement se fait en fonction donc des informations portées par les composantes c'est-à-dire selon la teneur en caroténoïdes, en anthocyanes et avec la première partie du red-edge. Les génotypes ayant des scores négatifs sur la composante PC1 (les génotypes situés à gauche) montrent un rendement potentiellement élevé sous des conditions optimales d'irrigation. A l'inverse, les génotypes avec des scores négatifs sur la composante PC1 ont des rendements potentiels plus faibles. Les

coordonnées sur la deuxième composante correspondent à la capacité des génotypes à répondre face à ce stress. En effet, un changement en teneur en caroténoïdes ou chlorophylle traduit une réponse face au stress (Pessarakli, 1999; Homayoun et al., 2011; Zarco-Tejada et al., 2002). Les génotypes ayant des scores positifs sur PC2 (F et D) possèdent beaucoup plus de caroténoïdes de manière générale contrairement aux deux génotypes I et C.

Terme interaction



(a) (b)

FIGURE 4.10 – Description du terme d'interaction :(a) Loadings de la première composante principale, (b) Loadings de la deuxième composante principale.

Les loadings des deux premières composantes du terme d'interaction sont illustrés figure 4.10. Sur la première composante (Fig. 4.10a), un creux est situé à 545 nm et une bosse est située à 672 nm. Un autre creux très important est situé à 729 nm. Une bosse apparaît de 800 nm à 922 nm. À partir 950 nm, les loadings sont proches de zéro. La première partie des loadings de la première composante (Fig. 4.10a) possède une forme similaire aux loadings de la première composante du terme génotype (Fig. 4.8a) qui exprime différentes teneurs en anthocyane et en chlorophylle. Cependant le pic situé dans le red-edge à 729 nm est légèrement décalé par rapport à la première composante du terme du génotype (Fig. 4.8a) et aucune cassure dans la pente n'est visible.

Les loadings de la deuxième composante (PC2) sont illustrés sur la figure 4.10b. Étonnamment, les loadings sont nuls de 400 à 700 nm. Une pente est visible de 750 nm à 1000 nm.

Cette seconde composante ne porte strictement aucune information sur la région spectrale visible et donc aucune information sur les pigments. Elle exprime ici des changements dans la structure interne des feuilles. En effet, des changements tels que des modifications dans la turgescence et l'espace intercellulaire peuvent être importants même en présence d'un stress hydrique très faible (Levitt and Zaken, 1975) À court terme, ce stress peut entraîner un flétrissement ou un enroulement de feuilles.

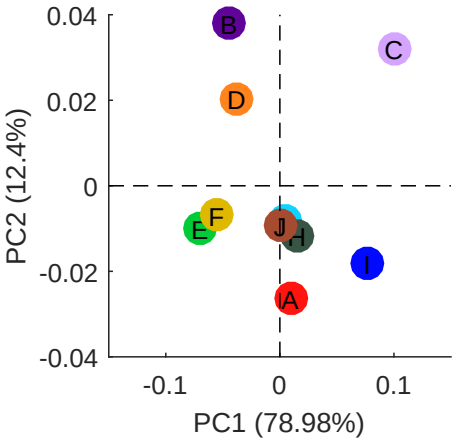


FIGURE 4.11 – Scores obtenus sur les deux premières composantes PC1, PC2 du terme d'interaction.

Les scores obtenus sur les composantes PC1 et PC2 par génotypes et selon le traitement sont présentés figure 4.11. Certains génotypes sont regroupés : les génotypes F et E ont des scores négatifs sur les deux axes ; les génotypes J/H/G ont des scores proches de zéros sur les deux axes. Sur la composante PC1, les génotypes E/F ont des scores négatifs. Ces génotypes qui ont une bonne teneur en chlorophylle et en anthocyane lorsqu'ils sont irrigués ont alors tendance à répondre différemment au stress. C'est-à-dire que les mêmes génotypes en condition non-irriguée ont une teneur en pigment beaucoup plus faible. Sur cette même composante, le génotype A présente peu de différences entre les traitements. Les traitements du génotype A ne se différencient pas selon la région spectrale liée aux anthocyanes et de la position du red-edge. Pour ce même génotype, le score observé sur la seconde composante PC2 est négatif mais de valeur absolue élevée. La différence entre les deux conditions d'irrigation concerne alors une différence de séchage de cellules internes et donc de flétrissement des feuilles. En suivant ce raisonnement, les génotypes B, D et C possèdent des scores positifs sur PC2. Ces génotypes ne possèdent pas de dégradation de structure interne. En revanche, le génotype C se dissocie nettement des génotypes B et D sur la composante PC1. Le génotype C agit alors à l'inverse des génotypes E et F, à savoir que la teneur en anthocyane et en caroténoïde est plus élevée dans la partie irriguée que dans la partie non-irriguée.

Pour chaque génotype, chaque observation peut produire un score par projection du spectre mesuré sur la composante étudiée. On peut ainsi représenter la variance des scores obtenus pour chaque génotype à travers des ellipses de confiance (Fig. 4.12). Les ellipses de confiance correspondent à un intervalle de confiance pour des réponses bivariées. Nous utilisons ici ces ellipses de confiance pour réaliser une comparaison de la variance des scores obtenus pour deux génotypes connus pour avoir des sensibilités très

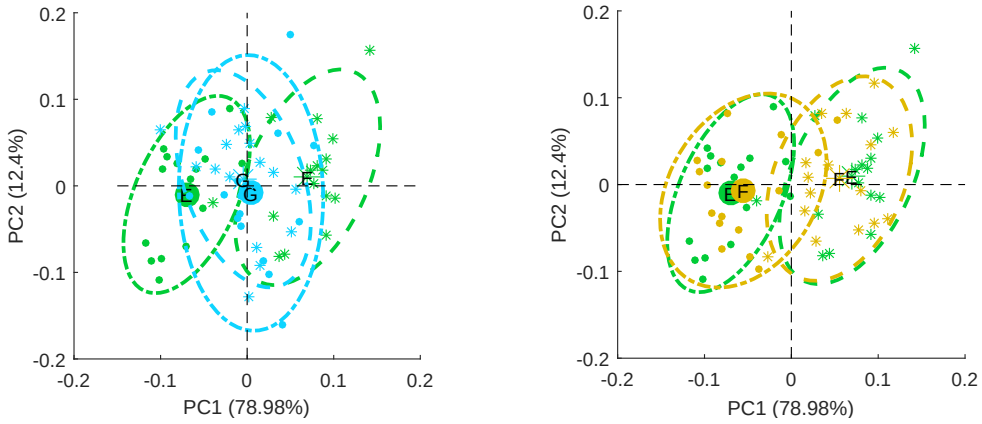


FIGURE 4.12 – Scores-plot et ellipses de confiance associées sur les deux premières composantes des génotypes :(a) E et G, (b) E et F.

différentes au stress hydrique, i.e. génotype E et G (Fig. 4.12a) puis pour deux génotypes connus pour avoir des comportements similaires, i.e. génotype E et F (Fig. 4.12b).

Les ellipses du génotype E pour les différents traitements sont bien dissociées alors que les ellipses du génotype G se chevauchent (Fig. 4.12a). De plus, la surface des ellipses du génotype G est supérieure à celle des ellipses du génotype E. Le génotype G présente donc une variance-intra élevée tant sur PC1 que sur PC2 avec des ellipses de grande surface qui ne se dissocient pas selon le traitement. Premièrement, la dissociation des ellipses de confiance (comme pour le génotype)E traduit ici une réponse précise du génotype vis-à-vis de la différence de traitement. Inversement un chevauchement traduit un manque d'adaptation vis-à-vis d'un stress comme c'est le cas pour le génotype G. Deuxièmement, la surface de l'ellipse correspond à l'hétérogénéité des réponses spectrales pour un traitement donné. Ces réponses spectrales correspondent ici à des changements de mécanismes mis en place face à un stress. Pour des surfaces grandes, le génotype réagit de façon hétérogène. À l'inverse, des ellipses de confiance de petites surfaces traduiraient ici une stabilité phénotypique (Becker and Leon, 1988). Les génotypes peuvent donc être aussi caractérisés/classés selon l'homogénéité de leur réponse spectrale et vis-à-vis de la dissociation face au traitement.

Les ellipses de confiance de deux génotypes tolérants (E et F) sont représentées sur la figure 4.12b . Comme dans le cas précédent, les ellipses entre les différents traitements du génotype E sont bien dissociées. Pour le génotype F, les ellipses entre les deux traitements se chevauchent très légèrement. On remarque que les deux génotypes se ressemblent très fortement aussi bien au niveau des barycentres que les ellipses de confiance. Le génotype F met alors en place des mécanismes face au stress comme le fait le génotype E. En effet, ces deux génotypes sont plutôt tolérants au stress hydrique. De cette étude, il

est possible de tirer un certain nombre d'indicateurs tels que la zone d'intersection des ellipses de confiance entre les deux traitements ou l'étendue de l'intervalle de confiance. L'hétérogénéité se traduit ici par la variance-intra décrite de manière représentative par l'ellipse de confiance (Worley et al., 2013).

Nous avons vu qu'il était possible de caractériser un certain nombre de traits phénotypiques sur la base des données spectrales. Les informations dépendent ici du facteur étudié. Nous allons donc dériver l'ensemble de ces résultats en traits phénotypiques pour les relier à des paramètres physiologique.

4.4 Recherche de traits phénotypiques

REP-ASCA permet de fournir par facteur des composantes formées de loadings. Pour chaque observation, un score peut être obtenu en projetant le spectre observé sur cette composante. Les scores générés de cette manière peuvent être utilisés directement comme traits phénotypiques. Il est possible de calculer un certain nombre de paramètres à partir de ces scores comme par exemple une différence entre scores par groupe de génotype ou des distances sur plusieurs composantes.

Dans cette partie, nous allons voir comment utiliser les scores comme traits et comment dériver un certain nombre d'indicateurs pour l'ensemble des facteurs. Pour le terme traitement, nous allons étudier les possibles relations existantes entre les scores moyens de chaque génotype ou les différences de scores par traitement avec le rendement. Pour les termes génotype et terme d'interaction génotype et traitement, nous allons étudier les scores moyens par génotypes.

4.4.1 Terme traitement

Des paramètres peuvent être dérivés à l'aide de la composante obtenue pour le terme traitement. Nous proposons les scores obtenus pour chaque génotype sur cette composante ainsi qu'une distance de score appelé d qui calcule pour chaque génotype la différence de scores par traitement. Les scores pour chaque génotype sont contenus dans une matrice $\mathbf{S}$ de dimension $\mathbb{R}^{g\times c}$ où g est le nombre de génotypes et c le nombre de composantes (égale à 1 pour le terme traitement). Avec $\mathbf{D}$, la matrice design disjonctive codant l'appartenance des observations aux génotypes, la matrice $\mathbf{L}$ de dimension $\mathbb{R}^{K\times c}$ correspondant aux loadings des composantes et la matrice $\mathbf{X}$ des mesures spectrales, la matrice des scores $\mathbf{S}$ est obtenue de la manière suivante :

$$\mathbf{S} = (\mathbf{D}^t\mathbf{D})^{-1}\mathbf{D}^t\mathbf{X}\mathbf{L} \tag{4.2}$$

La matrice de distance peut être obtenue de la manière suivante :

$$\mathbf{S}_{diff} = ((\mathbf{D}_1^t\mathbf{D}_1)^{-1}\mathbf{D}_1^t - (\mathbf{D}_2^t\mathbf{D}_2)^{-1}\mathbf{D}_2^t)\mathbf{XL} \tag{4.3}$$

Où les matrice design $\mathbf{D}_1$ et $\mathbf{D}_2$ codent les appartenances des génotypes pour respectivement le traitement irrigué et le traitement sec.

On peut alors étudier les corrélations existantes entre ces variables calculées $\mathbf{S}$ et $\mathbf{S}_{diff}$ et les variables agronomiques telles que le rendement.

Ces variables peuvent ainsi être reliées aux variables rendement et de pourcentage de pertes en rendement en étudiant les corrélations. La table 4.6 présente les corrélations obtenues entre $\mathbf{S}$, $\mathbf{S}_{diff}$ et la variable rendement.

TABLE 4.6 – Corrélation de la variable rendement avec les scores bruts du facteur traitement scores $\mathbf{S}$ et une différence de scores par génotype entre les deux traitements $\mathbf{S}_{diff}$.

Variable agronomique	$\mathbf{S}$	$\mathbf{S}_{diff}$
Rendement	-0.72	-0.66

Les corrélations obtenus pour ces variables de R=-0.72 et de -0.66 pour respectivement $\mathbf{S}$ et $\mathbf{S}_{diff}$. Les scores obtenus par génotype sont corrélés négativement à la variable rendement (R=-0.72). Cette corrélation est négative et confirme la constatation faite précédemment sur les scores (Fig. 4.7b). En effet, un score positif traduit un comportement de conditions non irriguées et donc avec potentiellement un rendement plus faible. Pour chaque génotype, la distance de scores entre les traitements est moins corrélée que les scores moyennés par génotype. Un score plutôt positif traduit un comportement moyen se rapprochant des comportements moyens en conditions non irriguées. Or, il est connu que le rendement est fortement réduit lorsque les plantes adoptent une stratégie de sauvegarde arrêtant la photosynthèse. Il n'est donc pas étonnant que le rendement soit fortement relié au score obtenu sur la première composante.

4.4.2 Terme génotype

Le terme génotype possède plusieurs composantes. Il est alors possible de calculer de nouveaux paramètres pouvant reliant ces composantes. Nous décidons d'illustrer cette démarche en représentant les scores obtenus sur les deux premières composantes en coordonnées polaires (Fig. 4.13). Un angle thêta et une distance à l'origine rhô sont alors obtenus.

La table 4.7 présente les corrélations obtenues entre variables calculées et la variable agronomique rendement (rendement moyen et pourcentage de pertes). On remarque

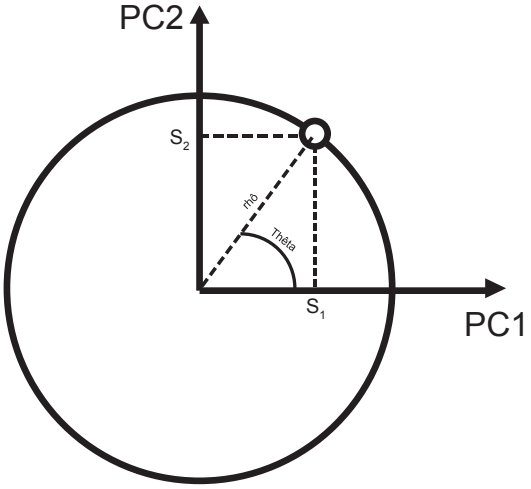


FIGURE 4.13 – Transformation des scores en coordonnées polaires avec rhô et thêta.

TABLE 4.7 – Corrélation des variables agronomiques avec les scores obtenus pour le terme génotype.

Variable agronomique	$\mathbf{S}_1$	$\mathbf{S}_2$	rhô	thêta
Rendement partie irriguée	-0.48	-0.59	-0.2	-0.42
Rendement partie non-irriguée	-0.34	0.03	-0.49	0.33
Rendement moyen	-0.51	-0.38	-0.38	-0.15
Pourcentage de perte	-0.24	-0.59	0.1	-0.62

que les scores obtenus sur la deuxième composante ($\mathbf{S}_2$) sont corrélés négativement (R=-0.59) au pourcentage de perte en rendement. Concernant les coordonnées polaires, rhô est négativement corrélé au rendement en condition non-irriguée (R=-0.49). La corrélation entre thêta et le pourcentage de perte est la plus élevée avec une valeur de R=-0.62. L'angle thêta semble être un bon critère pour classer les génotypes. En revanche, la distance à l'origine n'a pas d'influence.

Les scores sur la deuxième composante permettent de relier les génotypes à leurs tolérances au stress hydrique. Le calcul de nouveaux paramètres comme l'angle thêta permet d'améliorer cette caractérisation. La position des scores dans le plan formé des deux premières composantes permet d'apprécier les prédispositions d'un génotype pour être tolérant à un stress hydrique.

4.4.3 Terme interaction

Pour le terme d'interaction entre le génotype et le traitement, les variables utilisées comme traits sont les scores obtenus sur les deux premières composantes ainsi que leurs coordonnées polaires thêta et rhô.

TABLE 4.8 – Corrélation des variables agronomiques avec les scores obtenus pour le terme d'interaction génotype-traitement.

Variable agronomique	$\mathbf{S}_1$	$\mathbf{S}_2$	rhô	thêta
Rendement partie irriguée	0.69	0.17	0.02	0.35
Rendement partie non-irriguée	-0.2	0.71	0.27	0.57
Rendement moyen	0.41	0.47	0.14	0.53
Pourcentage de perte	0.81	-0.34	-0.23	-0.04

La table 4.8 montre les corrélations obtenues entre ces variables et le rendement. On constate que les scores $\mathbf{S}_1$ sont corrélés positivement à la variable rendement en condition irriguée avec un coefficient de corrélation de 0.69. Les scores $\mathbf{S}_2$ de la seconde composante sont corrélés positivement à la variable rendement en condition non irriguée avec un coefficient de corrélation de R = 0.71. Le pourcentage de perte est fortement corrélé (R = 0.81) aux scores obtenus sur la première composante. Rhô et thêta sont faiblement corrélés aux variables rendement.

Pour la variable rendement, le pourcentage de perte est corrélé fortement aux scores obtenus sur la première composante. Il est donc possible de classer la susceptibilité des génotypes selon les informations contenues sur la première composante. Des scores négatifs correspondent alors à des génotypes tolérants et des scores positifs à des génotypes sensibles au stress. Les loadings de la première composante contiennent des informations sur les pigments et le red-edge (Fig. 4.10a) tandis que la seconde composante exprime des variations de structure (Fig. 4.10b). Cela confirme qu'en condition de stress, les rendements obtenus peuvent se différencier à travers des différences de natures structurelles comme le flétrissement, l'assèchement ou la dégradation de structure cellulaire. En revanche, les variations de teneurs pigmentaires et du red-edge différencient alors les rendements sous conditions optimales d'irrigation ainsi que leur faculté à mobiliser le système photosynthétique face à un stress.

Pour la partie irriguée, les génotypes peuvent donc se différencier selon une différence en teneur de pigment (gamme spectrale visible). Pour la partie non irriguée, les génotypes se différencient selon les changements de structure interne des feuilles. La figure 4.14, reprend les scores $\mathbf{S}_1$ pour l'ensemble des génotypes (Fig. 4.11) en fonction du pourcentage de pertes. En effet, les génotypes B, D, E et F sont alors classés comme tolérants tandis que les génotypes J, C, G, H et I sont classés plutôt sensibles. On retrouve bien ce classement en reprenant la table donnée 4.3. Les génotypes ayant des scores négatifs possèdent tous des pourcentages de pertes inférieures à 11%. Tandis que les génotypes ayant des scores positifs sur cet axe possèdent des pourcentages de pertes supérieures à 20 %. Le génotype E est très atypique, car il est extrêmement tolérant avec une perte de rendement inférieure à 2 %. Le génotype ayant des scores proches des

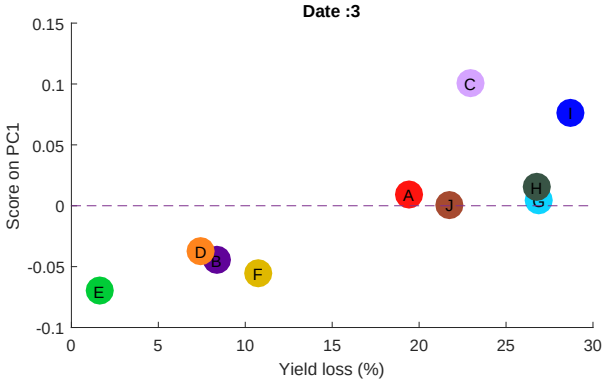


FIGURE 4.14 – Scores en fonction du pourcentage de pertes (Loss).

scores du génotype E est le génotype F. En revanche, les génotypes B et D se dissocient bien des génotypes E et F sur la deuxième composante principale. Ces deux couples se dissocient donc d'un point de vue de la structure des feuilles.

4.5 Conclusion du chapitre

Dans cette étude, nous nous sommes intéressés à l'intérêt d'une méthode telle que REP-ASCA pour décrire les réponses des génotypes vis-à-vis à un stress hydrique sur la base de données spectrales. Cette méthode a été testée sur un jeu de données comprenant un panel de génotypes ayant des sensibilités très différentes face au stress hydrique. Une seule date a été choisie pour étudier l'ensemble des facteurs dans un contexte de différence significative entre les deux conditions d'irrigation. Le terme de traitement permet ainsi de mettre en évidence les régions spectrales impactées par le changement d'irrigation.

Lorsque l'on s'intéresse au potentiel de rendement quel que soit le traitement, les génotypes peuvent se regrouper selon les composantes fournies par le terme génotype. Et enfin lorsque l'intérêt est de caractériser la réponse des génotypes face au stress, le terme d'interaction permet d'obtenir de nouveaux traits de phénotypage basés sur ces réponses spectrales. L'interprétation des loadings a permis dans cette étude de dissocier les génotypes sur les différentes stratégies mises en place face au stress. En cas de sécheresse légère, maintenir la croissance et le développement permet d'assurer un rendement. Cependant les génotypes dépourvus de l'un des mécanismes d'adaptation peuvent entraîner de graves dommages en termes de croissance, de développement et de rendement.

Bibliographie

Nathalie Vigneau. *Potentiel de l'imagerie hyperspectrale de proximité comme outil de phénotypage : application à la concentration en azote du blé*. PhD thesis, Montpellier, SupAgro, 2010. URL http://www.theses.fr/2010NSAM0026.

Llorenç Cabrera-Bosquet, José Crossa, Jarislav von Zitzewitz, María Dolors Serret, and José Luis Araus. High-throughput phenotyping and genomic selection : The frontiers of crop breeding converge. *Journal of integrative plant biology*, 54(5) :312–320, 2012.

José Luis Araus and Jill E. Cairns. Field high-throughput phenotyping : the new crop breeding frontier. *Trends in Plant Science*, 19(1) :52–61, January 2014. ISSN 13601385. doi : 10.1016/j.tplants.2013.09.008. URL http://linkinghub.elsevier.com/retrieve/pii/S1360138513001994.

Martin Ecarnot, Frédéric Compan, and Pierre Roumet. Assessing leaf nitrogen content and leaf mass per unit area of wheat in the field throughout plant cycle with a portable spectrometer. *Field Crops Research*, 140 :44–50, January 2013. ISSN 03784290. doi : 10.1016/j.fcr.2012.10.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0378429012003486.

G. R. Mahajan, R. N. Pandey, R. N. Sahoo, V. K. Gupta, S. C. Datta, and Dinesh Kumar. Monitoring nitrogen, phosphorus and sulphur in hybrid rice (Oryza sativa L.) using hyperspectral remote sensing. *Precision Agriculture*, December 2016. ISSN 1385-2256, 1573-1618. doi : 10.1007/s11119-016-9485-2. URL http://link.springer.com/10.1007/s11119-016-9485-2.

Nathalie Vigneau, Martin Ecarnot, Gilles Rabatel, and Pierre Roumet. Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat. *Field Crops Research*, 122(1) :25–31, April 2011. ISSN 03784290. doi : 10.1016/j.fcr.2011.02.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0378429011000451.

Anne-Katrin Mahlein, Ulrike Steiner, Christian Hillnhütter, Heinz-Wilhelm Dehne, and Erich-Christian Oerke. Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. *Plant methods*, 8(1) :1, 2012. URL http://plantmethods.biomedcentral.com/articles/10.1186/1746-4811-8-3.

Jan Behmann, Jörg Steinrücken, and Lutz Plümer. Detection of early plant stress responses in hyperspectral images. *ISPRS Journal of Photogrammetry and Re-*

mote Sensing, 93 :98–111, 2014. URL http://www.sciencedirect.com/science/article/pii/S092427161400094X.

Christoph Römer, Mirwaes Wahabzada, Agim Ballvora, Francisco Pinto, Micol Rossini, Cinzia Panigada, Jan Behmann, Jens Léon, Christian Thurau, Christian Bauckhage, and others. Early drought stress detection in cereals : simplex volume maximisation for hyperspectral image analysis. *Functional Plant Biology*, 39(11) :878–890, 2012. URL http://www.publish.csiro.au/?paper=FP12060.

Osval A. Montesinos-López, Abelardo Montesinos-López, José Crossa, Gustavo los Campos, Gregorio Alvarado, Mondal Suchismita, Jessica Rutkoski, Lorena González-Pérez, and Juan Burgueño. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods*, 13(1) :4, 2017. URL https://plantmethods.biomedcentral.com/articles/10.1186/s13007-016-0154-2.

Fernando M. Aguate, Samuel Trachsel, Lorena González Pérez, Juan Burgueño, José Crossa, Mónica Balzarini, David Gouache, Matthieu Bogard, and Gustavo de los Campos. Use of Hyperspectral Image Data Outperforms Vegetation Indices in Prediction of Maize Yield. *Crop Science*, 57(5) :2517, 2017. ISSN 0011-183X. doi : 10.2135/cropsci2017.01.0007. URL https://dl.sciencesocieties.org/publications/cs/abstracts/57/5/2517.

Nieves Aparicio, Dolors Villegas, Jaume Casadesus, José Luis Araus, and Conxita Royo. Spectral Vegetation Indices as Nondestructive Tools for Determining Durum Wheat Yield. *Agronomy Journal*, 92(1) :83, 2000. ISSN 1435-0645. doi : 10.2134/agronj2000.92183x. URL https://www.agronomy.org/publications/aj/abstracts/92/1/83.

M. A. Babar, M. P. Reynolds, M. van Ginkel, A. R. Klatt, W. R. Raun, and M. L. Stone. Spectral Reflectance to Estimate Genetic Variation for In-Season Biomass, Leaf Chlorophyll, and Canopy Temperature in Wheat. *Crop Science*, 46(3) :1046, 2006. ISSN 1435-0653. doi : 10.2135/cropsci2005.0211. URL https://www.crops.org/publications/cs/abstracts/46/3/1046.

E. Farshadfar and J. Sutka. Screening drought tolerance criteria in maize. *Acta Agronomica Hungarica*, 50(4) :411–416, 2002.

Willem Kruijer, Martin P. Boer, Marcos Malosetti, Pádraic J. Flood, Bas Engel, Rik Kooke, Joost J. B. Keurentjes, and Fred A. van Eeuwijk. Marker-Based Estimation of Heritability in Immortal Populations. *Genetics*, 199(2) :379–398, February 2015.

ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.114.167916. URL http://www.genetics.org/lookup/doi/10.1534/genetics.114.167916.

J. Bolanos and G. O. Edmeades. The importance of the anthesis-silking interval in breeding for drought tolerance in tropical maize. *Field Crops Research*, 48 (1) :65–80, 1996. URL http://www.sciencedirect.com/science/article/pii/0378429096000366.

Sylvie Parey, Farida Malek, Carine Laurent, and Didier Dacunha-Castelle. Trends and climate evolution : Statistical approach for very high temperatures in France. *Climatic Change*, 81(3-4) :331–352, March 2007. ISSN 0165-0009, 1573-1480. doi : 10.1007/s10584-006-9116-4. URL http://link.springer.com/10.1007/s10584-006-9116-4.

R. Chapuis, C. Delluc, R. Debeuf, F. Tardieu, and C. Welcker. Resiliences to water deficit in a phenotyping platform and in the field : How related are they in maize ? *European Journal of Agronomy*, 42 :59–67, October 2012. ISSN 11610301. doi : 10.1016/j.eja.2011.12.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1161030111001444.

Muhammad Aslam, Muhammad Amir Maqbool, and Rahime Cengiz. *Drought Stress in Maize (Zea mays L.)*. SpringerBriefs in Agriculture. Springer International Publishing, Cham, 2015. ISBN 978-3-319-25440-1 978-3-319-25442-5. doi : 10.1007/978-3-319-25442-5. URL http://link.springer.com/10.1007/978-3-319-25442-5.

R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, 43(5) :772–777, July 1989. ISSN 0003-7028, 1943-3530. doi : 10.1366/0003702894202201. URL http://journals.sagepub.com/doi/10.1366/0003702894202201.

Yiming Bi, Liang Tang, Peng Shan, Qiong Xie, Yong Hu, Silong Peng, Jie Tan, and Changwen Li. Interference correction by extracting the information of interference dominant regions : Application to near-infrared spectra. *Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy*, 129 :542–550, August 2014. ISSN 13861425. doi : 10.1016/j.saa.2014.03.080. URL https://linkinghub.elsevier.com/retrieve/pii/S1386142514004910.

Yiming Bi, Kailong Yuan, Weiqiang Xiao, Jizhong Wu, Chunyun Shi, Jun Xia, Guohai Chu, Guangxin Zhang, and Guojun Zhou. A local pre-processing method for

near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica Chimica Acta*, 909 :30–40, February 2016. ISSN 00032670. doi : 10.1016/j.aca.2016.01.010. URL https://linkinghub.elsevier.com/retrieve/pii/S000326701630054X.

Frank Boochs, K. Dockter, Gunther Kupfer, and W. Kuhbauch. Red edge shift as vitality indicator for plants. In *Proceedings of the 16th Congress of the International Society for Photogrammetry and Remote Sensing, held in Kyoto, Japan, on*, pages 1–10. Committee of the 16th International Congress for Photogrammetry and Remote . . . , 1988.

H. W. Gausman and W. A. Allen. Optical parameters of leaves of 30 plant species. *Plant Physiology*, 52(1) :57–62, 1973.

H. W. Gausman, W. A. Allen, and D. E. Escobar. Refractive index of plant cell walls. *Applied optics*, 13(1) :109–111, 1974.

Steven J. Crafts-Brandner and Michael E. Salvucci. Sensitivity of Photosynthesis in a C4 Plant, Maize, to Heat Stress. *Plant Physiology*, 129(4) :1773–1780, August 2002. ISSN 0032-0889, 1532-2548. doi : 10.1104/pp.002170. URL http://www.plantphysiol.org/lookup/doi/10.1104/pp.002170.

Hartmut K. Lichtenthaler and Joseph A. Miehé. Fluorescence imaging as a diagnostic tool for plant stress. *Trends in Plant Science*, 2(8) :316–320, August 1997. ISSN 13601385. doi : 10.1016/S1360-1385(97)89954-2. URL https://linkinghub.elsevier.com/retrieve/pii/S1360138597899542.

B. G. Osborne, T. Fearn, and P. T. Hindle. *Practical NIR spectroscopy with applications in food and beverage analysis*. Addison-Wesley Longman Ltd, Harlow UK, 1993. ISBN 978-0-582-09946-3. URL http://discovery.ucl.ac.uk/267166/.

Susan L. Ustin, David Riaño, and E. Raymond Hunt. Estimating canopy water content from spectroscopy. *Israel Journal of Plant Sciences*, 60(1) :9–23, December 2012. ISSN 0792-9978. doi : 10.1560/IJPS.60.1-2.9. URL http://www.sciencefromisrael.com/openurl.asp?genre=article&id=doi:10.1560/IJPS.60.1-2.9.

Tatjana Kavar, Marko Maras, Marjetka Kidrič, Jelka Šuštar Vozlič, and Vladimir Meglič. Identification of genes involved in the response of leaves of Phaseolus vulgaris to drought stress. *Molecular Breeding*, 21(2) :159–172, February 2008. ISSN 1380-3743, 1572-9788. doi : 10.1007/s11032-007-9116-8. URL http://link.springer.com/10.1007/s11032-007-9116-8.

M Chaves, João Maroco, and Joao Pereira. *Understanding plant responses to drought - From genes to the whole plant*, volume 30. January 2003. doi : 10.1071/FP02076.

M. M. Chaves, J. Flexas, and C. Pinheiro. Photosynthesis under drought and salt stress : regulation mechanisms from whole plant to cell. *Annals of botany*, 103(4) :551–560, 2009.

Barbara Demmig-Adams. Carotenoids and photoprotection in plants : A role for the xanthophyll zeaxanthin. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1020(1) : 1–24, October 1990. ISSN 0005-2728. doi : 10.1016/0005-2728(90)90088-L. URL http://www.sciencedirect.com/science/article/pii/000527289090088L.

Anatoly A. Gitelson, Yoav Zur, Olga B. Chivkunova, and Mark N. Merzlyak. Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy. *Photochemistry and Photobiology*, 75(3) :272–281, 2002.

Mohammad Pessarakli, editor. *Handbook of plant and crop stress*. Books in soils, plants, and the environment. M. Dekker, New York, 2nd ed., rev. and expanded edition, 1999. ISBN 978-0-8247-1948-7.

Houman Homayoun, Morteza Sam Daliri, and Parisa Mehrabi. Effect of drought stress on leaf chlorophyll in corn cultivars (Zea mays). *Middle-East J Sci Res*, 9(3) :418–420, 2011. URL http://idosi.org/mejsr/mejsr9(3)11/19.pdf.

Pablo J. Zarco-Tejada, John R. Miller, G. H. Mohammed, Thomas L. Noland, and P. H. Sampson. Vegetation stress detection through chlorophyll a+ b estimation and fluorescence effects on hyperspectral imagery. *Journal of environmental quality*, 31(5) : 1433–1441, 2002.

J. Levitt and R. Ben Zaken. Effects of Small Water Stresses on Cell Turgor and Intercellular Space. *Physiologia Plantarum*, 34(3) :273–279, July 1975. ISSN 0031-9317, 1399-3054. doi : 10.1111/j.1399-3054.1975.tb03835.x. URL http://doi.wiley.com/10.1111/j.1399-3054.1975.tb03835.x.

H. C. Becker and J. Leon. Stability Analysis in Plant Breeding. *Plant Breeding*, 101(1) : 1–23, August 1988. ISSN 0179-9541, 1439-0523. doi : 10.1111/j.1439-0523.1988.tb00261.x. URL http://doi.wiley.com/10.1111/j.1439-0523.1988.tb00261.x.

Bradley Worley, Steven Halouska, and Robert Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2) :102–104,

February 2013. ISSN 0003-2697. doi : 10.1016/j.ab.2012.10.011. URL http://www.sciencedirect.com/science/article/pii/S0003269712005131.

Application des proxis obtenus : influence la date de mesure et de l'environnement étudié

Chapitre 5

Sommaire

5.1 Introduction **107**

5.2 Influence de la date de mesure sur les proxis **109**

5.2.1 Application sur les autres dates de mesures 109

Terme traitement 109

Terme génotype 110

Terme interaction 113

5.3 Influence de l'environnement sur les proxis **114**

5.3.1 Description de la campagne expérimentale 2017 115

5.3.2 Application des composantes sur un autre environnement . . 116

Terme traitement 116

Terme génotype 117

Terme interaction 118

5.4 Conclusion du chapitre **120**

5.1 Introduction

Le chapitre précédent a montré que les données spectrales peuvent être utilisées comme proxi pour accéder à des informations précises sur la réponse des génotypes face à un stress hydrique. Ces résultats ont été obtenus pour une date discriminante quant à l'intensité du stress hydrique. Avec REP-ASCA, nous avons pu mettre en évidence des régions spectrales directement liées au pourcentage de perte de rendement. Or, le spectre de réflectance change au cours des différents stades phénologiques (Baret et al., 1987; Pena-Barragan et al., 2006). Le stress s'est produit à un stade phénologique spécifique correspondant au remplissage du grain. Un stress survenu à un autre stade n'aurait

certainement pas eu le même impact sur les réponses physiologiques (Okono; Aslam et al., 2015). Ces réponses peuvent se traduire par des changements dans la structure cellulaire ou de nature bio-chimique (pigments, présence d'eau). Le spectre de réflectance est alors susceptible de varier selon le moment et la nature du stress subi.

Dans ce chapitre, nous étudierons la répétabilité de ces résultats. Pour cela, nous étudierons l'effet date d'observation pour un même environnement. Puis, nous vérifierons si les proxis proposés peuvent être utilisés d'un environnement à l'autre. Il est possible d'utiliser la méthode REP-ASCA pour étudier l'effet date ou environnement. Or dans un contexte de mise en place opérationnelle, nous détaillons la démarche pour utiliser les proxis obtenus sur d'autres dates de mesures. Cette étude est réalisée dans la limite des données disponibles à savoir avec les deux campagnes expérimentales menées en 2017 et 2018.

Dans un premier temps, nous étudierons la transposabilité des informations obtenues à une date précise de mesure. La date utilisée est celle étudiée dans le chapitre précédent. Elle correspond au moment où le différentiel entre les deux traitements est maximal.

Dans un second temps, nous regarderons l'influence de l'effet environnement sur ces proxis utilisés. Cette partie représente la plus grande difficulté. En effet, d'un environnement à l'autre, les conditions agronomiques et pédoclimatiques sont différentes. Pour cela, nous utiliserons les proxis obtenus sur l'ensemble des observations effectuées d'une autre année. Non seulement la sévérité du stress est affectée, mais d'autres stress biotiques ou abiotiques peuvent être présents, non pris en compte dans notre modèle.

Nous répondrons aux questions suivantes :

— Comment utiliser les composantes décrites dans REP-ASCA dans une démarche opérationnelle ?

— Est-ce que le proxi est robuste vis-à-vis de la date de mesure ?

— Est-ce que le proxi est applicable dans un autre environnement pour décrire la réponse des génotypes au stress hydrique ?

Dans ce chapitre, nous étudierons l'impact de la date de mesure et de l'environnement sur les proxis correspondant aux différents termes étudiés (traitement, génotype et interaction). Les objectifs sont de vérifier que les composantes obtenues permettent de décrire les observations sur l'ensemble des dates de mesures. Il s'agit alors de vérifier que la composante du terme traitement permet de différencier les micro-parcelles irriguées des micro-parcelles non-irriguées ou que les composantes des deux autres termes permettent de caractériser respectivement les génotypes ou leurs réponses face à un stress hydrique.

5.2 Influence de la date de mesure sur les proxis

Sur la campagne expérimentale 2018 détaillée en section 4.2, les spectres ont été acquis à 6 dates différentes. En réalisant les acquisitions sur les mêmes parcelles à une date ultérieure, les spectres sont susceptibles de changer. Cela est dû à l'évolution du stade phénologique ou à un changement physiologique en réponse à un stress de la plante entre deux dates d'acquisitions. De plus, ces réponses diffèrent selon le génotype. Tous ces éléments impactent la mesure spectrale.

L'objectif de cette partie est de pouvoir caractériser les observations, quelle que soit la date de mesure. Dans notre cas, seules les premières composantes des termes traitement et du terme d'interaction sont utilisées comme proxis (Fig. 4.7a et 4.10a). Pour le terme génotype, les deux composantes sont utilisées (Fig. 4.8a, 4.8b).

Les observations réalisées aux autres dates sont alors projetées sur ces différentes composantes. Ces projections permettent d'obtenir des scores qui représentent les nouvelles coordonnées des données spectrales initiales sur ces composantes. Ces scores sont par la suite reliés aux pourcentages de pertes en rendement entre les deux conditions étudiées à savoir non-irriguée/irriguée. Ce pourcentage de pertes permet en effet de classer les génotypes sur leurs tolérances face à un stress hydrique.

5.2.1 Application sur les autres dates de mesures

Lorsque l'on utilise des modèles de prédiction, les données à prédire sont centrées par rapport aux données du jeu d'étalonnage. Pour une mise en place opérationnelle, l'objectif ici est d'obtenir des indicateurs invariants à l'effet date. Or, nous savons que des variations spectrales peuvent apparaître entre dates de mesures. Ces variations spectrales sont alors fortement liées au stade phénologique. Nous avons donc fait le choix de centrer les observations par rapport à la date. A une date donnée, nous retirons des observations le spectre moyen obtenu à cette date avant de projeter les observations sur cette composante. Centrer de cette manière revient à dissocier l'effet date comme dans une analyse de variance de type REP-ASCA. En faisant cela, les observations ne portent plus l'information sur la date de mesure. Les observations centrées sont ensuite projetées sur les composantes principales pour fournir des scores.

Terme traitement

Pour le terme traitement, une seule composante principale est obtenue. Cela représente un cas simple d'analyse où l'objectif est de dissocier les parcelles irriguées des parcelles non irriguées sur les autres dates.

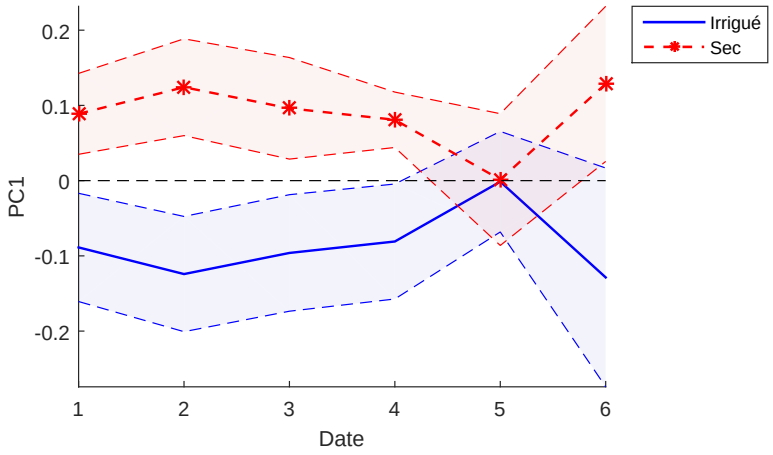


FIGURE 5.1 – Evolution des scores moyens sur la composante du terme traitement en fonction des dates d'acquisition.

Les scores moyens et les écarts-types par traitement sont obtenus en figure 5.1. A la date 5, les scores moyens par traitement se confondent et sont très proches de zéro. A cette même date, les écarts-types se chevauchent. Pour les autres dates, nous remarquons que les scores moyens de la partie non irriguée sont positifs et ceux de la partie irriguée sont négatifs. Les écarts-types pour ces dates ne se chevauchent pas.

A la date 5, la réserve hydrique de la parcelle irriguée devient faible (Tab. 4.4). De plus, cette date correspond à un stade de développement avancé. Il est donc possible que la différenciation entre partie irriguée et partie sèche soit difficile. Pour les autres dates, la classification correspond à l'étude faite au chapitre précédent (Section 4.3.4), à savoir que les micro-parcelles non irriguées ont majoritairement des scores positifs et les micro-parcelles irriguées des scores négatifs.

En appliquant directement la composante du terme traitement sur toutes les dates, les observations peuvent se différencier par traitement. Nous obtenons donc un proxi qui peut être utilisé à d'autres dates de mesure sans que la date ait une influence. Il est possible que ce proxi ne permette pas de mettre en évidence cette différenciation à certaines dates comme à la date 5. Cela est sans doute dû à un état de sénescence avancé, ne permettant plus de différencier les deux traitements. Ces éléments peuvent induire un changement dans les spectres.

Terme génotype

Pour le terme génotype, il est plus compliqué de détailler tous les génotypes. Nous avons fait le choix de montrer seulement deux couples qui se dissocient en termes de valeurs des scores sur ces deux composantes. Chaque couple est formé de deux génotypes ayant des scores proches et donc des comportements similaires. Les génotypes A et G

forment le couple de génotypes situés dans la partie coin inférieur gauche (scores négatifs sur les deux axes). Le deuxième couple formé des génotypes E et H est situé dans la partie supérieure droite (scores positifs sur les deux axes).

Des scores moyens par génotypes sont donc obtenus pour l'ensemble des dates (Fig. 5.2). Nous remarquons que les scores pour le couple E et H sur les deux premières composantes sont majoritairement positifs et donc visibles dans le coin supérieur droit. A la première date et à la dernière date, les génotypes de ce couple se distinguent très nettement : Les scores du génotype E sont proches de 0 sur la deuxième composante tandis que les scores du génotype H sont proches de zéro sur la première composante. Pour le second couple constitué des génotypes A et G, ces génotypes sont proches aux dates 3,4 et 5 mais sont très éloignés pour les autres dates. Les scores du génotype G possèdent des scores toujours négatifs sur la deuxième composante.

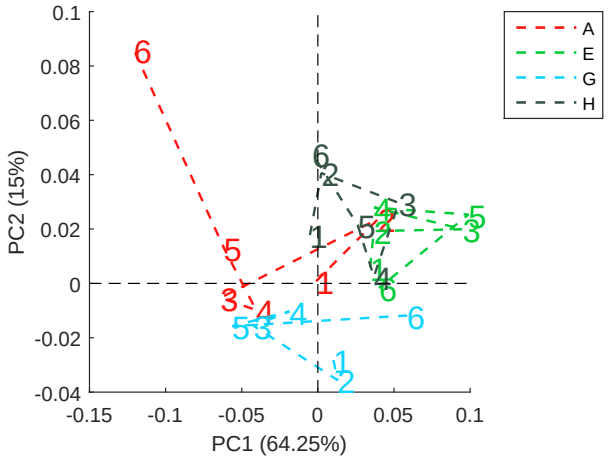


FIGURE 5.2 – Évolution des scores obtenus sur les deux premières composantes du terme génotype aux différentes dates pour les génotypes A, E, G et H. La numérotation indique les dates d'acquisition.

Pour le génotype A, les scores sur les deux composantes varient énormément d'une date à l'autre tandis que ceux du génotype E varient très peu. Étant très tolérant face au stress hydrique (Tab. 4.3), le génotype E semble être stable sur l'ensemble des dates contrairement au génotype A. Sur la première composante, les scores peuvent être négativement reliés au rendement moyen (Tab. 4.7) par la différence de teneurs en chlorophylle et en eau (Fig. 4.8a).

En s'aidant de l'interprétation des loadings de la deuxième composante (Fig. 4.8b), nous pouvons dire que les génotypes E et H ont des teneurs élevées en caroténoïde sur l'ensemble des dates. Le génotype G a une teneur en caroténoïde plus faible pour l'ensemble des dates tandis que cette teneur pour le génotype A fluctue énormément selon la date. Pour l'ensemble des dates, les génotypes E et H ont donc une activité

photosynthétique faible. Les génotypes A et G ont une activité photosynthétique élevée sur les dates 3,4 et 5.

Aux dates 3, 4 et 5, les génotypes de chaque couple sont très proches. Pour les dates de début de développement (1 et 2) et la dernière date (6), les génotypes de chaque couple sont bien dissociés.

Il est normal que des changements de classement de génotype apparaissent alors à des dates correspondant aux stades de début ou de fin de développement. Si un suivi temporel de la culture est envisagé, le chemin parcouru pourrait être un nouvel indicateur. Un chemin parcouru faible peut traduire une stabilité dans le comportement. Il est intéressant de noter que le génotype A continue l'activité photosynthétique à une date même très avancée (score négatif élevé sur la première composante à la date 6). Cela peut être l'effet d'une teneur élevée en caroténoïde représentée par un score très élevé sur la deuxième composante. Cette constatation peut être reliée au rôle photoprotecteur des caroténoïdes pendant la période de sénescence (Merzlyak and Gitelson, 1995).

Il a été souligné dans le chapitre précédent qu'un indicateur appelé thêta (Tab. 4.7) pouvait être relié au pourcentage de pertes. L'ensemble des caractères soulignés par les deux composantes (Fig. 4.8a et 4.8b) montre une prédisposition pour agir face au stress hydrique.

TABLE 5.1 – Corrélations entre l'indicateur thêta et le pourcentage de perte en rendement.

Date	Corrélation
1	0.36
2	-0.56
3	-0.62
4	-0.63
5	-0.17
6	0.23

Les corrélations pour l'ensemble des dates entre l'angle thêta et le pourcentage de perte en rendement sont montrés sur la figure 5.1. Nous remarquons que les corrélations les plus élevées sont obtenues pour les dates 2, 3 et 4 avec des valeurs respectives de -0.56,-0.62 et -0.63. Les valeurs sont en revanche très faibles pour les autres dates. Pour la première date et la dernière date, les coefficients de corrélation sont positifs. La composante étant obtenue à la date 3 où le système photosynthétique est installé et le couvert végétal fermé, il n'est pas étonnant de voir des inversions des valeurs de corrélation aux dates de début et de fin de stades de développement. On retrouve bien que ce coefficient est égal à -0.62 pour la date 3. Les corrélations sont également élevées aux dates voisines. L'angle thêta peut donc être un bon indicateur de la prédisposition

d'un génotype à répondre face à un stress hydrique, à condition de le calculer à des dates spécifiques. Ces dates doivent correspondre à un stade de développement similaire à la date 3.

Terme interaction

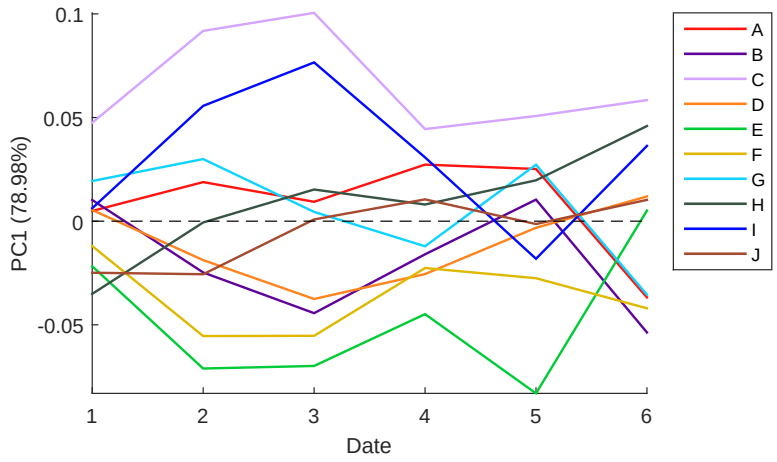


FIGURE 5.3 – Évolution des scores obtenus sur la première composante du terme interaction en fonction des dates d'acquisition et pour l'ensemble des génotypes.

Pour chaque date, des scores moyens par génotype sont obtenus par projection des observations sur la première composante du terme d'interaction (Fig. 5.3). Nous remarquons que ces scores évoluent selon les dates d'acquisitions. Les génotypes se dissocient ainsi de façon moins marquée sur la première et la dernière date. A ces dates, les scores sont relativement proches. Pour les autres dates, les génotypes E et C sont systématiquement situés aux extrémités avec des scores positifs pour le génotype C et des scores négatifs pour le génotype E.

A ces dates les génotypes C,I,A ont des scores toujours positifs tandis que les génotypes E, F et B ont des scores négatifs. Seuls les génotypes G, H et J changent de signes à ces dates.

On peut classer directement les génotypes selon les valeurs de leurs scores obtenues sur cette composante. Nous remarquons que ce classement est pratiquement inchangé sur les dates 1, 2,3 et 4. A la date 5, les scores sont tous plus proches de zéro hormis le génotype E qui se dissocie très bien des autres génotypes.

Le génotype E semble fortement atypique. En effet, c'est le génotype le plus tolérant avec très peu de différence entre les deux traitements (Tab. 4.1). L'écart maximal entre les génotypes est visible à la date 3. Cela est attendu car c'est à cette date que le déficit VPD est élevé et la différence de tensions entre les deux traitements est maximale. Pour

les dates 2, 3 et 4, les génotypes les plus tolérants ont des scores négatifs et les plus sensibles ont des scores positifs.

La première date et la dernière date correspondent aux dates de début et de fin de développement. On s'attend à voir peu de différence à ces dates entre les deux conditions d'irrigation. La différence d'irrigation est très faible à la première date tandis que la dernière date correspond à la période de sénescence. L'information contenue dans les spectres ne permet plus de dissocier les deux traitements d'irrigation.

TABLE 5.2 – Coefficients de corrélation entre les scores par date (obtenus avec la première composante de la date 3) et le pourcentage de perte en rendement.

Date	Corrélation
1	0.18
2	0.75
3	0.81
4	0.77
5	0.62
6	0.39

Dans le précédent chapitre (Chapitre 4), nous avons relié les scores de la première composante du terme d'interaction au pourcentage de perte de rendement et ainsi de classer les génotypes selon leurs tolérances face au stress hydrique. Les corrélations des scores obtenus pour les autres dates sont données dans la table 5.2. Nous remarquons que ces corrélations sont élevées pour les dates 2, 3, 4 et 5 avec des valeurs respectives de 0.75, 0.81, 0.77 et 0.62.

La corrélation la plus élevée est obtenue à la date 3 où le stress est le plus élevé. Le coefficient de corrélation est également élevé pour la date 4 où le différentiel de tensions dans le sol entre les deux traitements est plus faible (Tab. 4.4). La diminution de la contrainte hydrique agit lentement et donc à une date proche du stress maximal, il est normal d'obtenir des relations similaires.

Les valeurs des coefficients de corrélation confirment que la composante générée permet également de classer les variétés face au stress hydrique sur les autres dates à savoir les dates 2, 4 et 5. Toutefois, lorsque les mesures sont réalisées trop tardivement ou trop précocement, il est difficile de différencier les génotypes.

5.3 Influence de l'environnement sur les proxis

Nous avons vu que l'information obtenue à une date pouvait être utilisée pour d'autres dates de la même campagne. Dans cette partie, nous allons voir si ces mêmes composantes peuvent être utilisées pour les mesures réalisées dans un autre environnement. Pour

cela, nous utiliserons les données de la campagne expérimentale réalisée en 2017 sur un autre site de mesure. Dans un premier temps, nous décrirons les données agronomiques de cette campagne. Puis, dans un second temps, nous appliquerons les composantes obtenues dans le chapitre précédent (Chapitre 4) à ces données.

5.3.1 Description de la campagne expérimentale 2017

TABLE 5.3 – Déficit de pression de vapeur (VPD), tensiométrie (cbar) pour chaque traitement et différence de tensions entre les traitements pour l'ensemble des dates d'acquisitions de l'année 2017.

N°	Date	VPD (kPa)	P(non-irr)(cbar)	P(irr)(cbar)	Difference
1	19-Jun-17	1.53	5.83	5.83	x
2	19-Jul-17	0.89	13.08	11.41	1.67
3	02-Aug-17	0.95	73.42	50.08	23.34
4	17-Aug-17	0.80	43.92	26.01	17.91
5	01-Sep-17	0.31	2.92	1.42	1.5
6	28-Sep-17	0.47	1.42	1.42	x

La table 5.3 montre les valeurs de VPD et les valeurs tensiométriques de la campagne de 2017. A la première date, la valeur de VPD est de 1.53 kPa. Pour les autres dates, les valeurs de VPD sont inférieures à 1 kPa. Les tensions dans le sol sont toutes inférieures à 100 cbar. Le différentiel de tensions est faible et atteint un maximum à la date 3.

La tensiométrie et la VPD sont très faibles et il y a peu de différence de tensions entre les deux traitements. Les deux parcelles sont sous conditions optimales de réserve hydrique. Les parcelles ne sont pas stressées. Néanmoins, un différentiel léger apparaît entre les deux parcelles à la troisième date.

Pour chaque génotype, les valeurs de rendements pour les deux traitements et le pourcentage de perte entre les deux, sont données table 5.4. Les valeurs de rendement obtenues en condition irriguée sont élevées : 133.5 q/ha en moyenne. La valeur minimale est obtenue pour le génotype E avec une valeur de 117.6 q/ha. Les génotypes A, D et G ont des valeurs supérieures à 140 q/ha.

Pour la condition non-irriguée, les valeurs obtenues sont plus faibles que celles obtenues en condition irriguée mais restent élevées : 116.2 q/ha. Deux génotypes ont des valeurs faibles : les génotypes I et E avec des valeurs respectives de 99.5 q/ha et 103.6 q/ha. Pour l'ensemble des génotypes, le pourcentage de pertes varie entre 8 et 21% (avec une moyenne à 13%) Les pertes les plus élevées sont obtenues pour les génotypes C, I et A avec des valeurs respectives de 20.7, 20.5 et 17.8 q/ha. Les pertes les plus faibles sont obtenues pour les génotypes F, H avec des valeurs respectives de 8.4 et 8.9 q/ha.

TABLE 5.4 – Valeurs de rendements moyens par traitement et pourcentage de pertes (Loss) pour chacun des génotypes.

Genotype	Rendement		
	Irr	Non-irr	Loss (%)
A	141.6	116.3	17.8
B	131.9	118.4	10.2
C	138.5	110.1	20.5
D	140.8	125.1	11.1
E	117.6	103.6	11.9
F	128.5	117.7	8.4
G	142.6	127.4	10.6
H	134.7	122.8	8.9
I	125.4	99.5	20.7
J	133.6	120.9	9.5

Sur la base de ces éléments, il est difficile de classer les génotypes selon leurs tolérances au stress hydrique. En effet, le stress est très faible (Tab. 5.3) et les valeurs de rendement en condition non-irriguée sont trop élevées (Tab. 5.4) pour considérer cette modalité comme significativement stressée hydriquement. Par conséquent, les génotypes ne peuvent pas être caractérisés comme tolérants ou non tolérants sur cet essai. Le pourcentage de pertes traduit donc un effet bloc ou une variabilité génotypique de rendement due à de légères différences d'apport hydrique.

5.3.2 Application des composantes sur un autre environnement

Comme évoqué section précédente, il n'y a pas eu de stress hydrique important lors de la campagne expérimentale 2017. Nous allons tout de même appliquer les composantes principales des termes traitement, génotype et interaction présentées au chapitre précédent (Chapitre 4). En effet, même si le stress hydrique n'a pas eu un impact majeur sur le rendement, l'objectif est d'identifier des dates où un stress ponctuel a pu être présent, induisant une réponse physiologique des génotypes. Nous rappelons que ces composantes ont été obtenues à la date 3 de l'année 2018.

Terme traitement

La composante principale du terme traitement obtenue précédemment (Fig. 4.7a) est appliquée sur les observations réalisées pour l'ensemble des dates de la campagne expérimentale 2017.

La figure 5.4 montre les moyennes et les écarts-types des scores pour chaque traitement. Nous remarquons que les scores moyens de la partie irriguée sont négatifs pour

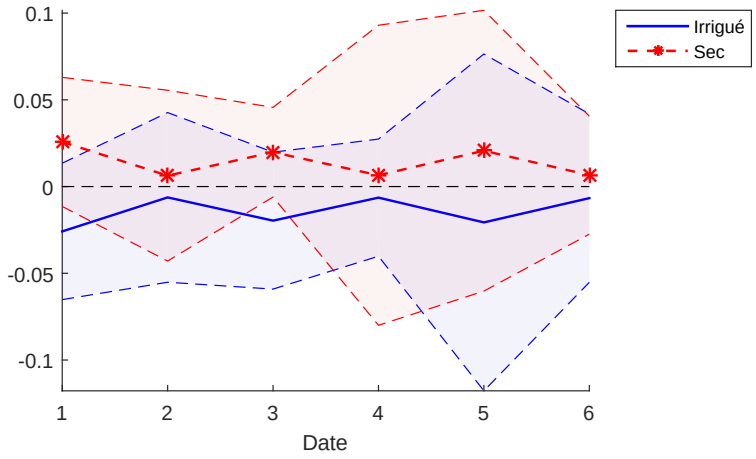


FIGURE 5.4 – Evolution des moyennes et des écarts-types des scores par traitement en fonction de la date.

l'ensemble des dates tandis qu'ils sont positifs pour la partie non-irriguée. Nous remarquons que les écarts-types sont élevés. Les valeurs se chevauchent entre les deux traitements pour l'ensemble des dates. A la date 3, ce chevauchement est plus faible.

On peut dissocier pour chacune des dates le traitement en moyennant les observations par traitement. En revanche, comme les écarts-types sont élevés, il est plus difficile de caractériser une observation isolée. Les écarts-types étant plus faibles à la date 3, il est plus facile de dissocier les observations pour les deux traitements.

L'écart très faible des scores moyens par traitement confirme que les conditions hydriques des deux traitements sont proches. Les traitements peuvent être cependant dissociés. La date 3 a été mise en évidence. En effet, les écarts-types obtenus sont plus faibles à cette date. On peut alors associer une partie des observations aux parcelles irriguées ou aux parcelles non-irriguées. Cette date correspond au maximum de différence de tensions dans le sol (Tab. 5.3).

Terme génotype

Dans le chapitre précédent, les scores calculés selon les deux premières composantes du terme génotype pour 4 génotypes ont été étudiés (Section 5.2.1).

Ces mêmes génotypes sont alors étudiés sur un autre environnement. Les scores obtenus sont montrés figure 5.5. Un regroupement de scores est visible dans la partie inférieure gauche de la figure. Peu de scores sont positifs sur les deux composantes simultanément. Seuls les génotypes H et A ont des scores positifs sur ces deux axes : aux dates 1, 3, 5 et 6 pour le génotype H ; aux dates 4 et 6 pour le génotype A. Nous remarquons également que les scores de trois génotypes (A, E et G) sont très proches à la date 5.

Comme cela a été évoqué lors de l'analyse des résultats de l'année 2018 (Section 5.2.1), des scores négatifs traduisent une forte activité photosynthétique et une faible teneur en caroténoïdes. A la cinquième date, les deux traitements ont des tensions extrêmement faibles. La disponibilité en eau entraîne une faible différentiation entre les génotypes A, E, et G. Des similarités avec les scores obtenus aux dates de l'année 2018 (Fig. 5.2) sont néanmoins visibles. En effet, le génotype H semble avoir une teneur élevée en caroténoïde pour une majorité des dates, ce qui est également le cas pour le génotype A à la dernière date. Comme pour l'environnement précédent, le génotype G ne possède aucun score positif sur la deuxième composante.

L'environnement semble avoir un impact important dans le regroupement des génotypes et particulièrement pour le génotype E. Lorsque les conditions hydriques sont optimales, les génotypes se ressemblent. Leur différentiation ne s'exprime pas car basée sur leur comportement face au stress hydrique.

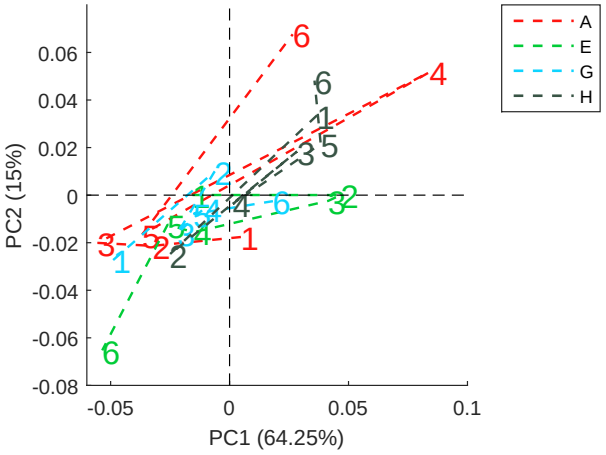


FIGURE 5.5 – Evolution des scores obtenus sur les deux premières composantes selon les dates pour un autre environnement.

Comme vu précédemment avec les scores observés sur ces 4 génotypes (Fig. 5.5), le regroupement des génotypes dépend fortement des dates de mesure et de l'intensité du stress hydrique à cette date. L'effet de l'environnement impacte alors fortement le classement des génotypes. De plus, lorsque l'environnement est sous conditions optimales d'irrigation comme à la date 5, les génotypes se dissocient très peu.

Terme interaction

Dans cette partie l'influence de l'environnement est étudiée sur les scores obtenus avec la première composante principale du terme d'interaction. Nous rappelons que sur cette composante, les génotypes avec des scores positifs sont classés comme sensibles

face au stress hydrique avec un pourcentage de perte en rendement élevé. A contrario, les génotypes avec des scores négatifs sont tolérants, le pourcentage de perte en rendement est faible.

Pour la campagne expérimentale de 2017, des scores sont calculés après la projection des observations sur la composante du terme d'interaction (Fig. 5.6). Nous remarquons que seuls les scores du génotype H ne changent pas de signes et sont toujours positifs. Pour les autres génotypes, les signes changent au cours des 6 dates de mesures et notamment à la date 5. En effet, le génotype B possède des scores négatifs sur l'ensemble des dates excepté à la date 5. Le génotype E a des scores négatifs pour les dates 2, 3 ,4 et 5. La courbe formée par les scores du génotype F semble suivre celle du génotype E.

Cette composante traduit une réponse face au stress hydrique. Puisqu'il n'y a pas eu de stress en 2017, il est normal que les valeurs obtenues soient très inférieures aux valeurs des scores de l'année 2018 (Fig. 5.3). Les changements de signe constatés à la date 5 sont liés au fait que les deux traitements ne présentent aucuns stress hydrique. Il n'est donc pas étonnement de voir des changements brusques de signes à cette date comme pour les génotypes I et B.

L'environnement a un fort impact dans le classement des génotypes pour cette composante principale. Cependant nous retrouvons certaines similarités avec les scores calculés aux dates de l'année 2018 (Fig. 5.3) à savoir : un génotype C possédant des scores élevés et positifs, des génotypes E et F proches (dates 2, 3 et 4) et le génotype B classé comme tolérant sur la majorité des dates.

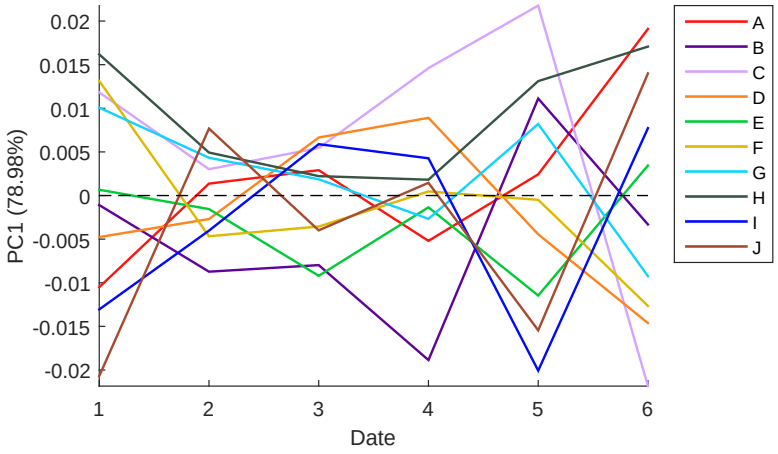


FIGURE 5.6 – Evolution des scores obtenus sur la première composante du terme d'interaction selon les dates pour un autre environnement

Si on se focalise sur la date 3, date à laquelle la réserve hydrique de la partie non-irriguée commence à se réduire légèrement, les scores peuvent être représentés en fonction des pourcentages de perte de rendement en 2018 (Fig. 5.7).

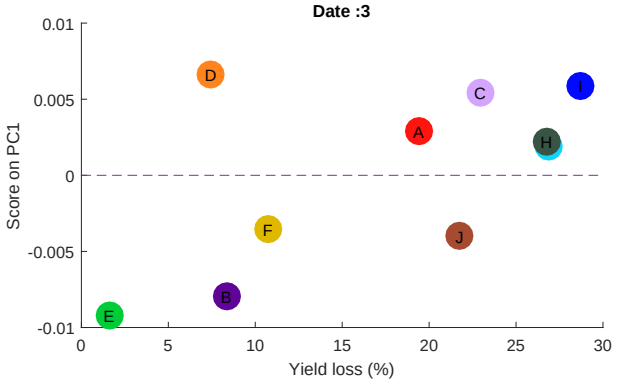


FIGURE 5.7 – Scores par génotype obtenus sur l'environnement de 2017 en fonction du pourcentage de pertes en rendement de l'environnement 2018.

Nous remarquons que les génotypes A,C,G,H et I ont des scores positifs et des pourcentages de pertes supérieurs à 20%. Les génotypes E, B et F ont des scores négatifs et de faibles pertes. Les génotypes D et J sont les seuls génotypes qui ne respectent pas cette constatation. Le génotype D a un score positif alors qu'il a eu très peu de pertes et le génotype J un score négatif alors qu'il a eu beaucoup de pertes.

Il est étonnant de retrouver de fortes similitudes avec le classement de l'année 2018 (Fig. 4.14) sur des observations effectuées sur un autre environnement présentant très peu de contraintes hydriques. Les génotypes E, B et F ont des scores négatifs sur les deux environnements et ont été classés tolérants au stress hydrique en 2018. Les génotypes C, G, H, A et I ont des scores positifs et ont été classés sensibles au stress hydrique en 2018. Pour les génotypes D et J, on peut émettre l'hypothèse que ceux-ci ne mettent pas en place ces réponses lorsqu'il n'y a pas de stress. La réponse spectrale peut alors ne pas présenter des caractéristiques de la composante utilisée.

Bien que le stress hydrique ait été très restreint, nous avons pu relier les scores obtenus à la date 3 avec le classement établi sur un environnement stressé hydriquement. Les scores obtenus sur cette composante sont corrélés positivement (R= 0.60) à leurs capacités à répondre au stress hydrique.

Ceci démontre que la méthode décrite peut être un bon moyen d'identifier les réponses différentielles des génotypes à un stress hydrique même dans une situation où ce stress est ponctuel, sans impact majeur sur le rendement.

5.4 Conclusion du chapitre

Nous avons vu qu'il était possible de calculer des scores à partir des composantes obtenues avec REP-ASCA. Ainsi, sur la base d'une date ou de plusieurs dates, nous pouvons obtenir des scores par observations.

Dans ce chapitre, nous avons étudié la possibilité de transposer les résultats issus de la REP-ASCA sur un autre stade phénologique ou un autre environnement. Les réponses spectrales évoluent tout au long du stade phénologique. Un stress subi à un moment donné peut potentiellement impacter les dates ultérieures.

Pour utiliser la démarche illustrée dans ce chapitre, il faut être vigilant avec le choix de la date. Nous avons vu que pour certaines dates, il était plus difficile de différencier les observations sur le critère du traitement lors des premières dates de mesures. Pour la majorité des dates, dans la mesure où celles-ci correspondaient à une période de stress hydrique, les conclusions tirées de l'information spectrale ont permis de classer les génotypes en fonction de leurs réponses faces aux stress.

Nous avons également appliqué des composantes obtenues par REP-ASCA à un autre environnement. L'environnement étudié n'a pas permis d'établir des conditions équivalentes de contraintes hydriques. Nous avons tout de même analysé les résultats pour l'ensemble des termes. Les scores obtenus pour la composante du terme traitement ont confirmés que les conditions hydriques étaient très similaires. Une date a pu être identifiée permettant de dissocier les deux traitements. Cette date correspond au différentiel maximal entre les deux conditions d'irrigation. Ce proxi permet alors d'observer de très légères différences de conditions d'irrigation. Les scores obtenus sur les deux composantes du terme génotype varient énormément selon la date de mesure.

La première composante principale du terme d'interaction a permis de générer des scores directement reliés au pourcentage de pertes en rendement. Cette composante est très robuste au changement de date de mesure. On a pu dissocier les génotypes tolérants des génotypes sensibles sur l'ensemble des dates de mesure hormis la première et la dernière date. On peut supposer qu'à ces dates correspondant respectivement aux stades végétatifs ou de sénescences, il est difficile d'obtenir des différences entre les variétés, soit parce que le stress hydrique ne s'est pas encore exprimé, soit parce qu'un stade de sénescence avancé a été atteint.

Cette même composante a été utilisée sur un autre environnement très peu stressé à une date présentant une très faible contrainte hydrique, nous avons pu rétablir le classement établi par l'environnement stressé avec les scores observés. Cette composante semble mieux s'exprimer lorsque les réserves hydriques sont limitées, mais reste néanmoins pertinente dans la caractérisation des variétés lorsque le stress est léger.

Il est alors envisageable de valider cette démarche sur d'autres environnements stressés hydriquement. Une campagne expérimentale est en cours avec notamment la validation d'un module couplant deux capteurs. Les composantes seront donc appliquées prochainement aux données d'imagerie permettant ainsi de valider la démarche.

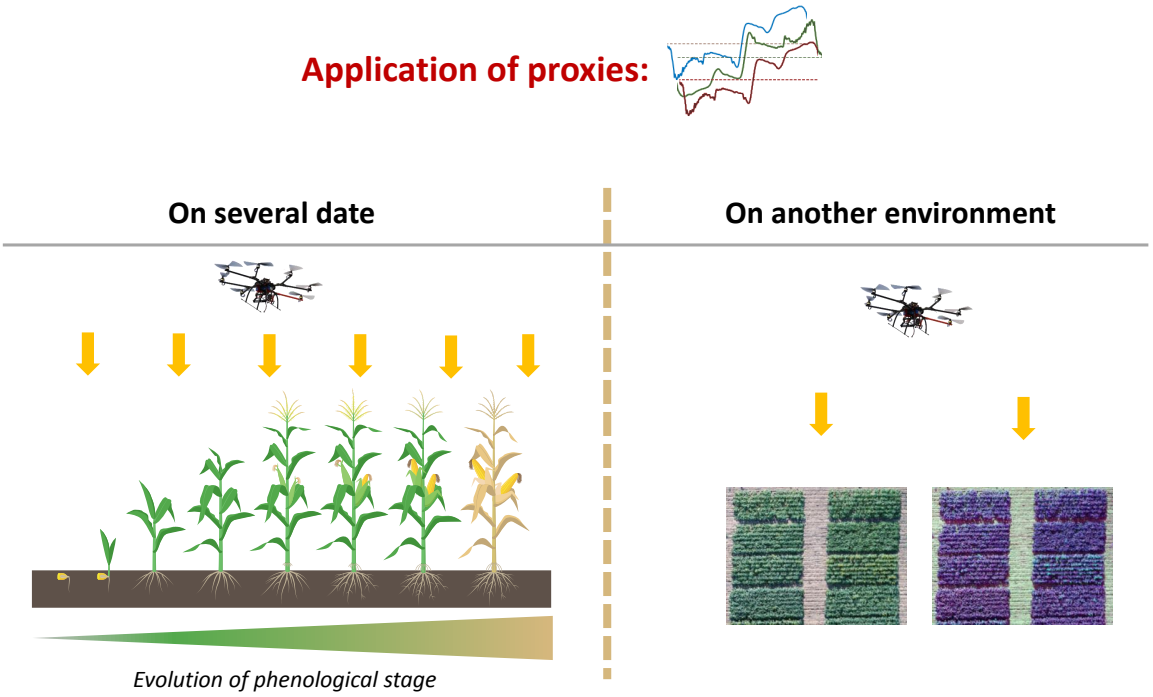


FIGURE 5.8 – Synthèse du chapitre 5 : Application des proxis obtenus par REP-ASCA sur d'autres dates et sur un autre environnement.

Bibliographie

F. Baret, I. Champion, G. Guyot, and A. Podaire. Monitoring wheat canopies with a high spectral resolution radiometer. *Remote Sensing of Environment*, 22(3) :367–378, August 1987. ISSN 0034-4257. doi : 10.1016/0034-4257(87)90089-7. URL http://www.sciencedirect.com/science/article/pii/0034425787900897.

J M Pena-Barragan, F Lopez-Granados, M Jurado-Exposito, and L Garcia-Torres. Spectral discrimination of Ridolfia segetum and sunflower as affected by phenological stage. *Weed Research*, 46(1) :10–21, February 2006. ISSN 0043-1737, 1365-3180. doi : 10.1111/j.1365-3180.2006.00488.x. URL http://doi.wiley.com/10.1111/j.1365-3180.2006.00488.x.

R. Okono. I. 1 Phenotyping drought-stressed crops : key concepts, issues and approaches.

Muhammad Aslam, Muhammad Amir Maqbool, and Rahime Cengiz. *Drought Stress in Maize (Zea mays L.)*. SpringerBriefs in Agriculture. Springer International Publishing, Cham, 2015. ISBN 978-3-319-25440-1 978-3-319-25442-5. doi : 10.1007/978-3-319-25442-5. URL http://link.springer.com/10.1007/978-3-319-25442-5.

M. N. Merzlyak and A. Gitelson. Why and what for the leaves are yellow in autumn ? On the interpretation of optical spectra of senescing leaves (Acer platanoides L.). *Journal of Plant Physiology*, 145(3) :315–320, 1995. ISSN 0176-1617. URL https://nebraska.pure.elsevier.com/en/publications/why-and-what-for-the-leaves-are-yellow-in-autumn-on-the-interpret.

Conclusion générale

Chapitre 6

Rappels

Cette thèse porte sur le couplage entre un capteur de haute résolution spectrale/faible résolution spatiale et un capteur à faible résolution spectrale et forte résolution spatiale pour la sélection variétale. Ce couplage est une solution efficace pour un transport sur drone. Cette association avec le drone permettrait de faire des acquisitions sur de grandes surfaces. Cette proposition représente un outil de phénotypage haut-débit à coût réduit.

Nous avons étudié comment reconstruire une image possédant à la fois une forte résolution spectrale et spatiale à partir de ce couplage. Un ensemble de méthodes appelées méthodes de pan-sharpening ont été testées via une approche par simulation. Une des méthodes s'est montrée particulièrement efficace : la méthode CNMF. En utilisant cette approche, nous avons développé deux méthodes permettant d'attribuer un spectre aux pixels non inclus dans l'image spectrale appelés pixels non-appariés. La méthode CNMF-KNN s'est montrée particulièrement performante. Elle s'appuie sur la similarité des valeurs de pixels appariés contenus dans le jeu d'entraînement. L'utilisation d'une image multispectrale à forte résolution spatiale améliore la qualité de reconstruction.

Le couplage de capteurs et la méthode de fusion proposée permettent ainsi de créer à bas coûts des solutions alternatives à l'imagerie hyperspectrale pour le phénotypage. L'utilisation d'une telle solution en extérieur peut induire des erreurs dues à un manque de répétabilité des mesures. En sélection variétale, il est courant d'avoir recours aux plans d'expériences et à l'analyse de variance. Cela permet de séparer les différentes sources de variabilités des mesures. Cependant l'erreur due au manque de répétabilité des mesures ne peut être dissociée des facteurs étudiés. De telles erreurs peuvent alors entacher les conclusions de l'analyse de variance. Une méthode appelée REP-ASCA a alors été développée pour réduire l'erreur de répétabilité. Cette méthode propose dans un premier temps d'étudier une série de répétition de mesure pour décrire cette erreur. Dans un second temps, alors que la caractérisation de cette erreur est réalisée dans

l'espace des observations appelé column-space, la réduction de cette erreur est effectuée dans l'espace des variables appelé row-space. Cette étape de réduction s'appuie sur la projection orthogonale.

Le potentiel d'une telle solution a été étudié sur deux campagnes expérimentales. L'objectif était de montrer la plus-value d'une forte résolution spectrale pour la discrimination des génotypes face à un stress hydrique. Pour ce faire, nous avons donc appliqué notre approche sur une date spécifique correspondant à un stress hydrique élevé. L'approche REP-ASCA a permis de générer des composantes par facteurs : traitement, génotype et interaction génotype/traitement. Ces composantes traduisent des changements dans les spectres qui peuvent être reliés à une modification de teneurs biochimiques ou à des changements de paramètres de structure. Il est possible de caractériser la réponse des génotypes face au stress notamment avec les composantes du terme d'interaction. Les scores obtenus peuvent alors être utilisés comme des indicateurs de stress.

Vers un outil opérationnel au champ

Le couplage entre un capteur de forte résolution spectrale/faible résolution spatiale et un capteur de faible résolution spectrale/forte résolution spatiale permet d'obtenir un outil pouvant être embarqué sur un drone. L'ensemble de la chaine de traitement protocole d'acquisition et méthodes d'analyse permet de générer de nouveaux traits phénotypiques. Ces traits phénotypiques correspondent aux composantes générées par REP-ASCA. Ces composantes permettent de discriminer les génotypes.

On peut ainsi dissocier deux modes opératoires pour l'utilisation d'un tel outil au champ : le premier mode pour l'acquisition de connaissances en utilisant REP-ASCA. Cette approche permet de générer des composantes et d'interpréter via les loadings, les éléments mis en place par des génotypes soumis à un stress. Le deuxième mode opératoire est d'utiliser simplement cet outil pour acquérir des données sur de nouvelles parcelles et d'appliquer directement les composantes obtenues lors du premier mode opératoire. Les scores générés pourraient alors être comparés à ceux obtenus sur panels de génotypes existants et connus ou être utilisés directement comme critères de sélection.

Perspectives scientifiques

Avec l'outil présenté précédemment, plusieurs campagnes expérimentales pourront être menées pour obtenir des composantes plus précises et dans des conditions contrastées de stress hydrique, ou encore pour des stress de natures différentes (maladie, stress chaud, etc.) ou selon de nouveaux critères de sélection (mélange variétal). Ainsi dans

des acquisitions futures, il sera d'autant plus facile de classer ou différencier de nouvelles variétés inconnues sur la base de ces nouvelles composantes. La conception d'un tel outil a été étudiée dans le cadre de la sélection variétale, il serait également possible d'utiliser cet outil pour d'autres applications en extérieur. Ces applications pourraient alors être de nature agronomique comme le suivi de culture, la détection de maladie ou de mauvaises herbes ou pour d'autres champs disciplinaires comme par exemple la classification d'espèces de couvert forestier.

Dans le cadre de cette thèse, la méthode REP-ASCA a été développée comme méthode d'analyse de variance pour des données multivariées et pour réduire l'erreur de répétabilité. La condition est d'ajouter au protocole de mesures une série de mesures destinée à caractériser cette erreur. Cette méthode peut ainsi être utilisée dans toute expérience incorporant des acquisitions de mesures multivariées pour réduire des erreurs dues au manque de répétabilité de la mesure. Cette méthode a d'ailleurs été validée sur des données spectrales de grains de café issues de la spectroscopie NIR. Ainsi, l'approche proposée peut être appliquée sur des données en provenance d'autres domaines produisant des données spectrales comme la métabolomique ou en chimie analytique.

Annexes

Chapitre A

Sommaire

A.1 Campagne expérimentale 2019 129

A.1.1 Introduction 129

A.1.2 Matériels et méthodes 130

Vol de drone 130

Traitement des données 130

A.1.3 Résultats et discussions 131

Description de la campagne expérimentale 2019 131

Résultats de REP-ASCA 132

A.1.4 Conclusion 133

A.2 Article sur la méthode REP-ASCA 134

A.1 Campagne expérimentale 2019

A.1.1 Introduction

Afin de mettre en évidence les principaux éléments de la thèse et de valider l'approche d'un couplage de capteurs, un module appelé AirStrip a été conçu et développé pendant la thèse. Ce module permet de réaliser des mesures spectrales conjointement à des prises d'images RGB (sans filtre infrarouge). Ainsi pour une image acquise, un spectre est associé à certains pixels de cette image. En étant embarqué par drone, cet outil fournit alors plusieurs images et plusieurs spectres.

Le traitement des données acquises par AirStrip reprend alors l'ensemble des méthodes développées par les chapitres. La première étape est d'utiliser les méthodes de pan-sharpening discutés au chapitre 2 pour reconstruire une image avec de fortes résolutions spectrale et spatiale. Cette image peut ensuite être utilisée pour une nouvelle

analyse de variance ou en appliquant directement les composantes issues de REP-ASCA.

De la même manière que le chapitre 5, nous présentons dans cette annexe l'application de la première composante du terme d'interaction obtenue au chapitre 4 sur les données acquises en 2019 par le module AirStrip.

A.1.2 Matériels et méthodes

Vol de drone

La campagne expérimentale de 2019 est située à Nérac et reprend le même plan d'expérience que les deux précédentes années avec 10 génotypes et 2 traitements d'irrigation. Un vol de drone a été effectué le 19/08/2019 en embarquant le module AirStrip en avec un plan de vol couvrant la parcelle (Fig. A.1). Le vol a duré 14 minutes générant plus de 1000 images et 1000 spectres associés.

FIGURE A.1 – Plan de vol du drone embarquant AirStrip pour la campagne de 2019

Traitement des données

Le traitement des données s'est effectué en deux étapes : la reconstruction de l'image hyperspectrale en utilisant l'algorithme CNMF-KNN (Section 2.5.1) et l'application de la composante du terme d'interaction (Fig. 4.10a) obtenue en 2018. La composante permet d'obtenir des scores par pixels.

Le spectromètre du module AirStrip (STS, Ocean Optics) possède une plage spectrale plus réduite. La première composante du terme d'interaction obtenue en 2018 (Fig. 4.10) a été a donc été adaptée (Fig. A.2). Pour que les scores soient moyennés par génotypes, les micro-parcelles ont été délimitées par des polygones.

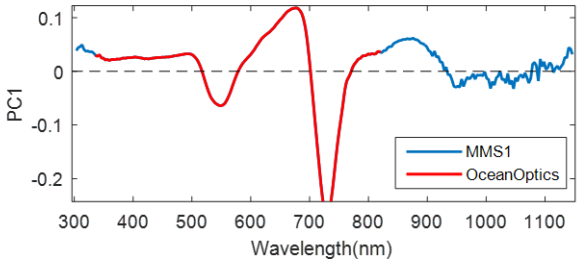


FIGURE A.2 – Première composante du terme d'interaction issue de la campagne 2018 adaptée au spectromètre Ocean Optics

A.1.3 Résultats et discussions

Description de la campagne expérimentale 2019

La table A.1 montre les valeurs de VPD et les valeurs tensiométriques à la date d'acquisition de 2019. La valeur de VPD est de 2.26 kPa à cette date. La tension dans le sol est de 239 cbar pour la partie irriguée et est de 175.75 cbar pour la partie non-irriguée. La tensiométrie et la VPD sont assez élevées pour les deux traitements. Cela correspond à un stress élevé.

TABLE A.1 – Déficit de pression de vapeur (VPD), tensiométrie (cbar) pour la campagne de l'année 2019.

Date	VPD (kPa)	P(non-irr)(cbar)	P(irr)(cbar)
19-Aou-19	2.26	239.00	173.75

Pour chaque génotype, les valeurs de rendements pour les deux traitements et le pourcentage de perte entre les deux, sont visibles table A.2. Dans des conditions optimales d'irrigations, les valeurs de rendement obtenues en condition irriguée sont assez basses avec une moyenne de 114.4 q/ha. La valeur minimale est obtenue pour le génotype D avec une valeur de 102.9 q/ha. Le génotype A possède la plus forte valeur de rendement en traitement irrigué. Les autres génotypes ont des valeurs entre 110 q/ha et 119 q/ha.

Pour le traitement non-irrigué, les plus fortes valeurs sont obtenues pour les génotypes F, G et D avec des valeurs supérieures à 100 q/ha. Les génotypes possédant les valeurs les plus faibles sont les génotypes J, E et A avec des valeurs de 73.5 q/ha, 88.1 q/ha et 91.1 q/ha.

Les pourcentages de pertes en rendement sont visibles en table A.2). Les génotypes A,E et J ont des valeurs très élevées en pourcentages de pertes avec respectivement, 25.1%, 25.9% et 36.9%. Les génotypes D,F et G ont les plus faibles valeurs en pourcentages de pertes en rendement avec respectivement 0.62%, 2.8% et 7.1%.

En comparant les valeurs des pourcentages de pertes en rendement en 2018 (table 4.3) et en 2017 (table 5.4), on retrouve difficilement des valeurs comparables pour les génotypes sensibles et tolérants. En effet, le génotype le plus tolérant du panel correspond au génotype E et possède en 2019 une forte valeur de pourcentage de pertes en rendement (25.9%). Selon ces valeurs, il est difficile d'établir précisément quels sont les génotypes sensibles et les génotypes tolérants.

TABLE A.2 – Valeurs de rendements moyens par traitement et pourcentage de pertes (Loss) pour chacun des génotypes.

Genotype	Rendement		
	Irr	Non-irr	Loss (%)
A	121.6	91.1	25.1
B	117.3	99.8	14.9
C	114.6	98.7	13.9
D	102.9	102.3	0.62
E	119.0	88.1	25.9
F	111.0	107.7	2.8
G	110.3	102.54	7.1
H	112.2	99.7	11.1
I	118.9	97.0	18.2
J	116.5	73.5	36.9

Résultats de REP-ASCA

La figure A.3 montre les scores obtenus par génotype en fonction du classement établi en 2018 du comportement des génotypes.

Les scores des génotypes E, D, B, H, J et G sont négatifs tandis que les scores des génotypes F, A, C, et J sont positifs. En comparant avec les scores obtenus en 2018 (Fig. 4.11), on retrouve des similitudes : les génotypes E, D et B, classés comme tolérants au stress hydrique ont des scores négatifs et les génotypes sensibles au stress hydrique (C, I et A) ont des scores positifs. En revanche, les génotypes G, J et H ne respectent pas ce classement établi en 2018.

Les valeurs de rendement obtenues à la récolte (table A.2) ne mettent pas en évidence le même classement que précédemment. Malgré l'influence de l'environnement (Section 5.3) et de la date de mesure (Section 5.2) sur les scores, les génotypes les plus tolérants E, D et B sont bien classés.

La figure A.4 montre un exemple de scores obtenus de deux micro-parcelles obtenus après la projection de l'image hyperspectrale sur la composante A.2. Les scores forment alors une image. L'échelle de couleur correspond aux valeurs des scores entre -0.15 et

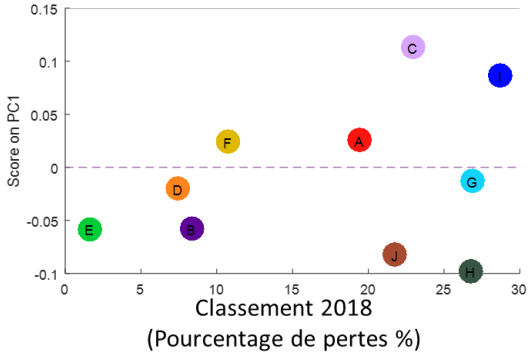


FIGURE A.3 – Scores obtenus pour la campagne 2019 par génotypes

+0.15. On remarque que les scores élevés correspondent aux pixels de sol. Concernant les pixels de végétation, les scores du génotype I (Fig. A.4a) se situent entre 0 et 0.1 tandis que ceux du génotype E (Fig. A.4b) se situent entre -0.1 et 0. Le génotype I correspond au génotype le plus sensible au stress hydrique tandis que le génotype E correspond au génotype le plus tolérant. On retrouve ces conclusions donc sur l'ensemble d'une micro-parcelle.

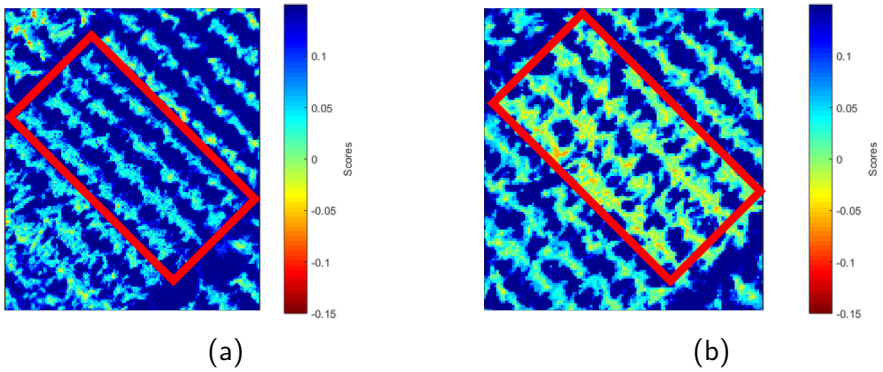


FIGURE A.4 – Illustration des scores obtenus sur les micro-parcelles correspondant aux génotypes (a) I et (b) E

AirStrip permet d'accéder à de l'information spatiale conjointement à de l'analyse d'information spectrale comme la distribution des scores au sein de la micro-parcelle.

A.1.4 Conclusion

Nous avons testé le module AirStrip à travers une campagne expérimentale de 2019 reprenant les mêmes génotypes étudiés précédemment. Ce module est une solution de couplage entre un capteur de faible résolution spatiale/forte résolution spectrale (spectromètre) et d'un capteur forte résolution spatiale/faible résolution spectrale (imageur RGB).

Ce module associé au drone répond aux contraintes de la sélection variétale et propose une solution de phénotypage rapide. Le traitement de données s'appuie sur la méthode CNMF-KNN proposée dans le chapitre 2 (Section 2.5.1) et permet de reconstruire une image avec une forte résolution spatiale et une forte résolution spectrale.

La composante obtenue en 2018 (Fig. 4.10a) a simplement été appliquée à cette image reconstruite produisant ainsi une image de scores. Ces scores peuvent être reliés aux différents comportements connus des génotypes. Les génotypes les plus tolérants au stress hydrique sont bien classés. Dans cette annexe, nous avons choisi d'exploiter l'information spectrale des micro-parcelles. Avec la solution AirStrip, l'information spatiale au sein de la micro-parcelle est également disponible. Nous pouvons alors envisager d'étudier conjointement ces informations de nature spatiale (surface de végétation, homogénéité au sein de la micro-parcelle, ...).

De nouvelles campagnes de mesures avec ce module sont prévues afin de le valider sur un panel de génotypes plus large et dans des environnements différents. Il faut alors inclure plusieurs dates d'acquisitions pour identifier la date idéale d'acquisition ou éventuellement étudier l'évolution temporelle de la signature spectrale.

A.2 Article sur la méthode REP-ASCA

Reduction of repeatability error for Analysis of variance-Simultaneous Component Analysis (REP-ASCA): Application to NIR spectroscopy on coffee sample

Maxime Ryckewaert[a,b], Nathalie Gorretta[b], Fabienne Henriot[a], Federico Marini[c], Jean-Michel Roger[b]

[a]*Limagrain Europe, Chappes, France*
[b]*ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, Montpellier, France*
[c]*Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy*

Abstract

A method to reduce repeatability error in multivariate data for Analysis of variance-Simultaneous Component Analysis (REP-ASCA) has been developed. This method proposes to adapt the acquisition protocol by adding a set containing repeated measures for describing repeatability error. Then, an orthogonal projection is performed in the row-space to reduce the repeatability error of the original dataset. Finally, ASCA is performed on the orthogonalized dataset. This method was evaluated on NIR spectral data of coffee beans. This study shows that the repeatability error due to physical variations between measurements can alter results of the analysis of variance. These effects are predominant in factors analysis and can be seen on spectra as constant or non-constant baselines. By reducing repeatability error

Email address: maxime.ryckewaert@limagrain.com (Maxime Ryckewaert)

with REP-ASCA, baselines are removed and factor analysis provides more information about chemical content of the factors of interest.

1. Introduction

When conducting measurements in experiments, measured values can differ from the true values of the samples. Measurement error is the difference between a measured quantity and its true value. For reliable measures, the measurement error must be relatively small compared to the variance of the measured samples [1].

The reproducibility error is the difference between two measurements of the same sample under two different conditions, for example, change of laboratory or operator. When conditions of acquisitions do not change, the reproducibility error is also called repeatability error [2]. In this paper, the repeatability error stands for differences observed between two measurements performed on the same physical sample.

In an experiment, the observed responses can contain repeatability error and this may affect the results of processing carried out on this data. This is especially true when we want to study the influence of independent variables, i.e., the so-called factors.

The use of design of experiments (DoE) allows defining the experiments to perform in an optimal way, in order to test the influence of the identified factors on the observed responses [3, 4]. The statistical analysis called the ANalysis Of Variance (ANOVA) of a dataset from an experimental design, was historically formalized in 1925 by Fischer with the first edition of [5]. The aim of this analysis is to split the total variance of the measured responses

into several sources of variances [6] However, the repeatability error may alter the results of ANOVA, even when the factors have a relevant influence on the response(s) of a studied system. Indeed, it may become more complicated to differentiate which part of the variance is explained by the factors and which part is due to the repeatability error.

In spectroscopy, measurement error may be due to both the variation of the measurement conditions (e.g. angle of view, reference, sensor temperature, etc) and the physical properties of the measured sample (e.g. particle size, roughness, physical texture, etc). In addition, a change in sample chemical composition modifies the light diffusion process. Measuring non-uniform samples adds a major difficulty compared to homogeneous powdered samples [7]. For example, variations in oil, water, fiber, protein, and mineral content of the same individual change the nature of the light-scattering medium. These chemical and physical variations can lead to unwanted variations in spectra. These variations can for example modify the spectra by adding them with baselines, which can be horizontal lines or slopes, even curves. When the changes are proportional to the intensity value, the effect is called multiplicative effect. These phenomena are well known in spectroscopy [8, 9] and are one of the sources of repeatability error.

Some sources (sensors, operator, etc.) of errors can be described as factors. If we consider these factors should have negligible effects on the total variance, we don't account for these factors in the analysis of variance and they are referred to as hidden. Some of these factors can be described as a random factor. In the case of univariate data, these random factors can be accounted for, using adapted methods such as Generalized Linear Mixed

Models (GLMMs) [10]. In the case of multivariate data, solutions lie in (1) reducing dimensionality or (2) modelling multiple responses (e.g., longitudinal analysis by using GLMMs; [11]). However, GLMMs are unsuitable in most cases, especially with spectral data [12].

For multivariate data, chemometric methods can be used to take advantage of the DoE [13]. Among those methods, the Analysis of Variance - Simultaneous Component Analysis (ASCA) [14] is the most used, but the repeatability error is not taken into account in ASCA. This method systematically averages the repetitions of measurements on an individual. Averaging observations helps to reduce the repeatability error but requires increasing the number of measurements. Climaco-Pinto et al. [15] propose a method consisting in reducing the residual variance after each step of the ANOVA-PCA. This method makes the separation of factors clearer by reducing within-variance. This method does not focus on the repeatability error.

In this study, a method named Reduction of the error of repeatability-Analysis of variance-Simultaneous Component Analysis (REP-ASCA) is proposed to reduce repeatability error by adding additionnal dataset. REP-ASCA is evaluated on NIR spectral data collected on coffee beans.

The objectives of this paper are: (1) to suggest a method (REP-ASCA) to reduce the repeatability error (2) to apply REP-ASCA to real data from NIR spectroscopy (3) to study the influence of the number of measurement in the additionnal set on the reduction of the repeatability error.

2. Theory

2.1. ASCA

Multivariate data can be represented as a matrix $\mathbf{X}$ of size (N, P) where N is the number of observations and P the number of the measured variables. The analysis of variance establishes a model in order to decompose the matrix $\mathbf{X}$ into different sources of variation. By doing so, relationships between the factors and the measured variables can be studied.

The ASCA method includes of three main steps: The first one consists of partitioning the observation matrix $\mathbf{X}$ into matrices associated with the effect of the studied factors and their interactions. For example, in a DoE involving two studied factors A and B, $\mathbf{X}$ can be written according to the ASCA approach as follows:

$$\mathbf{X} = \mu + \mathbf{X}_A + \mathbf{X}_B + \mathbf{X}_{AB} + \mathbf{E} \tag{1}$$

With μ the mean of all observations, $\mathbf{X}_A$ and $\mathbf{X}_B$ the terms associated respectively to the main effects of the factors A and B, $\mathbf{X}_{AB}$, the interaction term between A and B and $\mathbf{E}$ the residuals. Matrices $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_{AB}$ contain identical replicates of the average spectra collected at each level of the particular factor (A, B and AB, respectively). For example, for a given level of the factor A, all the rows of X_A corresponding to this level will contain the same spectrum.

For a complete and balanced design, factors A, B and AB are independent of one another [16]. More specifically, matrices $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_{AB}$ are all orthogonal to one another [17]. Thus: $var(\mathbf{X}) = var(\mathbf{X}_A) + var(\mathbf{X}_B) + var(\mathbf{X}_{AB}) + var(\mathbf{E})$. The term $var(\mathbf{E})$ contains the within-level variance and

the other terms the between-level variance for the corresponding factor. In practice, the variances of the factors are replaced by the sum of the squares of the elements of the associated matrix. For example, with $\|\mathbf{X}_A\|^2$ corresponding to the square Frobenius' norm of $\mathbf{X}_A$, the variance of $\mathbf{X}_A$ is replaced by [18, 17, 19]:

$$\mathrm{SSQ}(\mathbf{X}_A) = \|\mathbf{X}_A\|^2 \tag{2}$$

The second step of ASCA tests if factors have a significant effect on the total variance. To do that, a permutation test is performed and a p-value is estimated by comparing variance value on the original data with the distribution of variance obtained by permutations. The permutation test can be performed for data which does not necessarily respect the conditions of normality [12]. For a factor i and its associated matrix $\mathbf{X}_i$, the p-value is calculated according to the following equation [20]:

$$\text{p-value}(\mathbf{X}_i) = \frac{\mathrm{nbr}(\mathrm{SSQ}(\mathbf{X}_{i,perm} \geq \mathrm{SSQ}(X_i))}{k} \tag{3}$$

Where nbr() calculates the number of occurrences, k is the number of permutations and $\mathbf{X}_{i,perm}$ the matrix obtained after a random row permutations. The p-value is then calculated by counting the number of cases where the variance of the studied factor is lower than the variances resulting from the permutations. Indeed, by doing so, the effect of the studied factor is compared to its distribution under the null hypothesis as estimated by the permutations.

For the last step, a Simultaneous Component Analysis (SCA) [21, 22] is performed on all the terms of the matrices belonging to equation 1. Thus,

for a given factor i, the decomposition of the corresponding matrix $\mathbf{X}_i$ is written:

$$\mathbf{X}_i = \mathbf{T}_i \mathbf{P}_i^t \tag{4}$$

Where $\mathbf{T}_i$ and $\mathbf{P}_i$ correspond respectively to the scores and the loadings of the principal components. $\mathbf{T}_i$ and $\mathbf{P}_i$ are respectively matrices of size (N,L) and (P,L) with L is the number of levels of factor i minus one.

This analysis reduces the representation space. These loadings define a subspace spanned by $\mathbf{X}_i$ and highlight the spectral directions related to the factors studied. The scores are the new coordinates of the observations on these principal components.

2.2. Repeatability error

When a sample measurement is repeated several times, measurements are never identical. Observed variation between measurements is due to the repeatability error. When measurements are performed on different samples, variance differences are due to the variance between samples but also contain repeatability error. The greater the repeatability error, the greater the within-variance increases at the expense of between-variance. In fact, the variance of the factor becomes smaller relatively to the total variance. This tends to decrease the significance of the factors.

Some of the error can be due to a random phenomenon whose average is null. In ASCA, the repeatability error is reduced because repeated measures are systematically averaged over every sample. As a consequence, when the repeatability error is high, one solution is to reduce its variance by increasing

the number of observations. As a result, the number of acquisitions increases considerably. In addition, it is impossible to link the repeatability error to a specific factor in ASCA. Indeed, it is a nested factor [23]. The solution proposed in this study is to use an additional set of repeated measures to take away the repeatability error of the observations before analysis of variance. [24]

2.3. Proposed method: REP-ASCA

The method proposed in this study consists in identifying the spectral directions responsible for the repeatability error and removing them from $\mathbf{X}$. By definition, $\mathbf{X}$ can be represented in two dual spaces: the column-space of dimension R^N spanned by the observations and the row-space of dimension R^P spanned by the variables. In ASCA, the decomposition part of the variance is performed in the column-space. Equation 1 can be written as a series of orthogonal projections [17]:

$$\mathbf{X} = \mathbf{M}_1\mathbf{X} + \mathbf{M}_A\mathbf{X} + \mathbf{M}_B\mathbf{X} + \mathbf{M}_{AB}\mathbf{X} + \mathbf{E} \tag{5}$$

Where each matrix $\mathbf{M}_i$ is an orthogonal projection matrix in the column-space (R^N) defined by the following equation:

$$\mathbf{M}_i = \mathbf{D}_i(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t \tag{6}$$

Where $\mathbf{D}_i$ called the dummy-matrix or design matrix is a binary matrix encoding the class belonging of the factor i.

The variance of repeatability error can be included into all factors. This variance cannot be expressed into a unique factor and therefore represented

by an orthogonal projection. This error cannot be reduced in the column-space.

The spectral directions of the repeatability error can be described in the row-space through the space W spanned by the repeatability error. Let $\mathbf{X}_s$ be a matrix containing the repetitions of measurements performed on a set of samples representative of $\mathbf{X}$. Each repeated set is centered and all the centered sets are stored in a matrix $\mathbf{W}$. The matrix $\mathbf{W}$ contains only the effects related to the repetition of measurements. A Principal Component Analysis (PCA) is performed on $\mathbf{W}$ and provides loadings $\mathbf{P}$ which define the repeatability error subspace. The matrix $\mathbf{X}$ is then projected orthogonally to $\mathbf{P}$ using the following formula:

$$\mathbf{X}_{\perp} = \mathbf{X}(\mathbf{I} - \mathbf{P}\mathbf{P}^t) \tag{7}$$

ASCA is then performed on $\mathbf{X}_{\perp}$ instead of $\mathbf{X}$. This method contains two parameters to adjust: the number k of dimensions removed by the projection, i.e. the number of rows of $\mathbf{P}$ and the minimum number of spectra required in $\mathbf{W}$, i.e. the number of rows of $\mathbf{W}$.

The setting of the parameter k can be done according to a set of criteria. First, the scree plot of the PCA eigenvalues can be examined, and the value of k that corresponds to a net break is the selected. Another criterion is to study the evolution of the ratio between-variance/total-variance for the design factors and identify a break, a plateau. When the data are signals as it is the case with spectral data, studying shape of loadings (break, baseline) can provide information about the nature of repeatability errors and thus allow to adjust the k parameter accordingly. The REP-ASCA method assumes

that the repeatability error is structured, i.e. a large part of this error has a subspace of limited size, which can be removed by orthogonal projection into the row-space. When the number k is too high, the loadings become noisy.

When building $\mathbf{W}$, that is to say acquiring $\mathbf{X}_s$, it is necessary to check that enough repeated measures are available to describe the repeatability error. This can be verified by monitoring the evolution of the structure of the subspace spanned by $\mathbf{W}$. Let $\mathbf{J}_n$ be the inertia matrix of $\mathbf{W}$ when it contains n spectra. The RV coefficient [25, 26] between $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ is calculated for each addition of elements. The evolution of this coefficient as a function of n reflects the evolution of the subspace structure generated by $\mathbf{W}$. In practice, the filling of $\mathbf{W}$ is stopped when this coefficient no longer changes.

3. Materials

3.1. Experimental data

The REP-ASCA was tested on data resulting from near infrared spectroscopy of coffee beans [18]. The aim of the experiment was to study species and roasting times influences on spectra. Spectra were recorded over a spectral range of 1000 nm to 2500 nm with a resolution of 1.2 nm and transformed in pseudo-absorbance ($-\log 1/R$). Initially, ten geographical origins were compared representing a total of 800 spectra. For this study, only one origin (80 spectra) was selected.

Two species of coffee (Robusta and Arabica) were roasted during four different periods (0mn, 25mn, 50mn, 75mn) at a constant temperature of 180 °C. Seven successive spectral measurements were collected on each sam-

ple. This DoE yielded a matrix $\mathbf{X}$ of 56 spectra (2 species x 7 spectral measurement x 4 duration times). In addition, 3 repetitions of measurements have been made on each of the 8 samples leading to 24 spectra in the matrix $\mathbf{X}_s$.

3.2. REP-ASCA model

The matrix $\mathbf{X}$ was projected orthogonally to the k first loadings extracted from X_s yielding $\mathbf{X}_\perp$. The value of k was varied from 0 (classical ASCA) to 12. In accordance with ASCA modelling, $\mathbf{X}_\perp$ was decomposed into the following terms:

$$\mathbf{X}_\perp = \mu + \mathbf{X}_{species} + \mathbf{X}_{time} + \mathbf{X}_{speciesXtime} + \mathbf{E} \tag{8}$$

With: $\mathbf{X}_{species}$ and $\mathbf{X}_{time}$ factors related respectively to species and roasting duration time. $\mathbf{X}_{speciesXtime}$ is the interaction term and $\mathbf{E}$ the residuals.

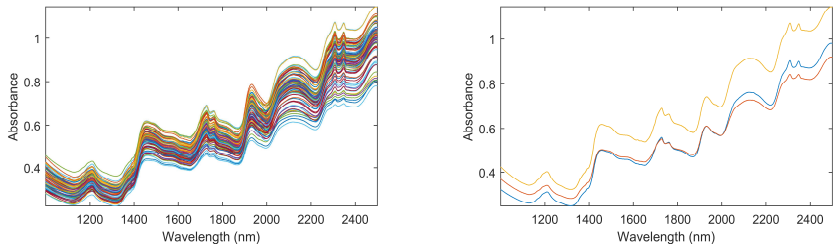
4. Results and discussion

4.1. Preliminary analysis of spectra

:

The $\mathbf{X}$ spectra are shown in 1a. We can observe some characteristic peaks of coffee spectra according to the literature [27, 28, 29]. Positions and characteristics of these peaks are presented in Table 1:

However, some differences between spectra can be observed. There is a shift in the baseline. This shift increases with the wavelength. This is frequently observed in NIR spectroscopy when the spectra are acquired in



(a) (b)

Figure 1: (a) All spectra forming by $\mathbf{X}$, (b) 3 spectra measured on one sample

Table 1: Position and peak characteristic of coffee spectra

Wavelength (nm)	Bond vibration	Assignment
1212	C−H stretching 2nd overtone	Caffeine
1450	C−H stretching 1st overtone	H_2O and Cellulose
1725/1765	C-H stretching 1st overtone CH_2	Fatty acids
1940	O−H stretching + O−H deformation	H_2O
2080-2150	C=O (1st overtone) and O–H (combination)	Protein/Fatty acids/Carbohydra
2309/2350	C–H + C=C (combination)	Carbohydrates : Cellulose

reflectance [30, 31, 9, 32]. This phenomenon is due to differences in the light scattering.

Figure 1 b shows three spectra of $\mathbf{X}_s$ measured on the same sample and illustrating the repeatability error. These spectra present the same kind of variation observed on $\mathbf{X}$ spectra (Fig. 1a). A vertical translation as a shift in a constant baseline appears between spectra at all wavelengths. This is because each spectrum was measured on a different area of the same sample. In each area, the reflectance intensity varies according to the geometric configuration of the measured grains in relation to the sensor. These aspects

are also observed over the whole databases $\mathbf{X}$ and $\mathbf{X}_s$.

The baselines also vary in slopes (see Fig. 1b). This is due to a variation of particle size and surface roughness of the product [30]. It can therefore be assumed that the internal structure of coffee beans varies within the same sample.

A more detailed analysis of 1b shows three distinct spectral regions for baseline variations:

- Between 1000 nm and 1400 nm, a negative slope is observed, which varies for each repetition.
- Between 1400 nm and 2000 nm, baselines are increasing, but appear to have identical slopes to ones observed at other spectral ranges.
- Between 2000 and 2500, baselines are increasing, but appear to have different slopes. It can be hypothesized that all of these baseline variations are due to several scattering regimes (e.g., Mie and Rayleigh). These scattering regimes are governed by the relationship between particle size and light wavelength [7].

All these baselines deform the information contained in the spectra. These undesired effects should be removed to better describe the factors.

4.2. ASCA on $\mathbf{X}$

The Explained variance and the p-value for each factor are obtained from ASCA on $\mathbf{X}$ and summarized in Table 2. We can then observe that the largest source of variance is explained by the residuals. 82.01% of the total variance could not be related to a studied factor. 5.93% of the variance is

Table 2: Explained variances and p-values of the factors

Factor	Explained Variance (%)	p-value
$\mathbf{X}_{species}$	5.93	0.056
$\mathbf{X}_{time}$	8.61	0.128
$\mathbf{X}_{speciesXtime}$	3.45	0.548
$\mathbf{E}$	82.01	

explained by the species, 8.61% of the total variance is explained by the roasting duration time and 3.45% by the interaction term.

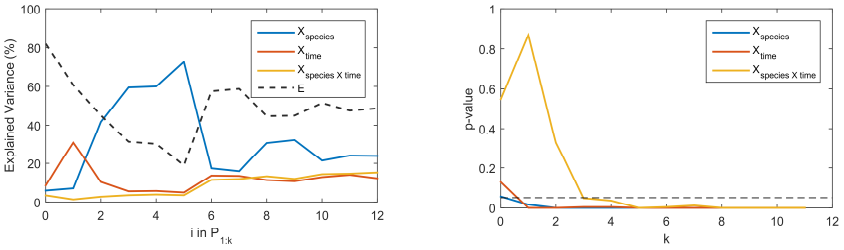
In addition, the p-values obtained after the permutation test, show that none of the factors is statistically significant at 5%. These results can be explained in part by a large part of variance not explained by the model. The factor variances are too low in comparison to the residual variance. Here, the case study shows that the repeatability error is so large that none of the identified factors are expressed. Within-variance of repeated measures are more important than variances between-level of factors. It is then very difficult in this particular case to draw conclusions with ASCA.

4.3. REP-ASCA

4.3.1. Selection of the number k of component removed

The parameter k, which corresponds to the number of components of $\mathbf{W}$ to remove, must be set. For this purpose, consequences on ASCA performed on $\mathbf{X}_{\perp}$in terms of explained variances and p-values are studied according to k.

Fig. 2a shows, that part of variance carried by $\mathbf{E}$ decreases up to the



(a) (b)

Figure 2: Evolution of: (a) explained variances by factors, (b) p-value for each factor ; according to the projection number.

5th component. Then, it subsequently increases. Inversely, the term $\mathbf{X}_{species}$ explains more and more variance and that up to the 5th component. The effects related to the repeatability error hide the expression of the term $\mathbf{X}_{species}$ to the 5th component. For the $\mathbf{X}_{time}$ term, projecting orthogonally to the 1st component increases the variance part from 8.61% to 30.81%. This large increase shows that the 1st component of $\mathbf{W}$ mainly masked variance due to roasting time. Nevertheless, the variance explained by this term decreases thereafter. This means that from the second component, the repeatability error is very much related to roasting. For the interaction term $\mathbf{X}_{speciesXtime}$ explained variance is very small. It increases very slowly as the repeatability error is removed.

By reducing the variance due to the repeatability error, the variances of the factors can be modified in different proportions. Information is extracted in different proportions from the different factors. The significativity of each factor can then potentially change. The evolution of p-value (Fig. 2b) shows that the repeatability error masks the effects of the factors. By removing the

information carried by the first component of $\mathbf{W}$ (i.e., projection orthogonally to the 1st component of the repeatability error), the $\mathbf{X}_{time}$ and $\mathbf{X}_{species}$ factors become both significant, in agreement with what reported in the original study [18]. The interaction term becomes significant after a projection orthogonal to the first three components.

4.3.2. Analysis of the repeatability error

Let's examine in detail the results of REP-ASCA for k=5. The sources of repeatability can be explored through the analysis of the loadings of the first 5 component (Fig 3).

Fig. 3a shows the percentages of explained variance by component. The first component explains 91.6% of the total variance. The loadings of the first component (Fig. 3b) are all positive. This indicates a systematic variation across all variables. The shape of these loadings is the same as the average spectrum of the individuals of $\mathbf{X}_s$. This reflects the vertical translation observed in (Fig. 1a).

The second component explains 6.53% (Fig. 3a) of the total variance. Surprisingly, the loadings of the second component (Fig. 3c) are similar to the loadings of the first component (Fig. 3b). However, slight differences can be observed: first, the loadings of the second component are centered on zero. This indicates an overall slope variation. Secondly, these loadings show a zero slope between 1000 nm and 1400 nm, whereas it is negative on the loadings of the first component of this same spectral range. As a result, the deflection is not homogeneous over all wavelengths.

The third component explains 1.13% (Fig. 3a) of the total variance. The loadings of the third component (Fig. 3d) show two slopes: one between 1000 nm and 1400 nm and another between 2000 nm and 2500 nm. This corresponds to different scattering regimes related to the size of the scattering media in respect to the wavelength. This scattering phenomenon is finer than what we saw before. This phenomenon results in a non-constant curving baseline effect. The fourth component explains 0.2% of the total variance (Fig. 3a). On the loadings (Fig. 3e), a first slope is visible between 1000 nm and 1300 nm and a second between 1900 nm and 2300 nm. Unlike previous components, slopes are less visible. This component is therefore less related to physical phenomena. This gives information related to the chemistry of the

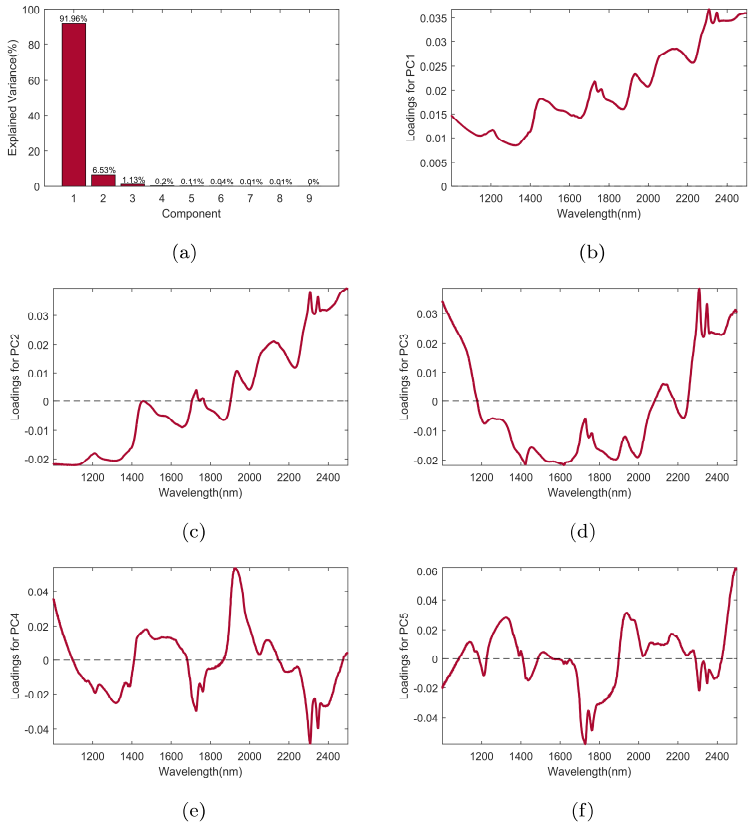


Figure 3: PCA performed on W: (a) Part of variances explained by principal components (b) Loadings on PC1, (c) Loadings on PC2 (d) Loadings on PC3, (e) Loadings on PC4, (f) Loadings on PC5.

medium. A characteristic peak at 1940 nm is related to the presence of water [33] as a result of the combination of the O-H stretch band and the O-H_2 torsion band. Two peaks are located at 1710 nm and at 2307 nm respectively linked to lipids and proteins content. The fifth component explains 0.11% of the total variance (Fig. 3a). As shown in (Fig. 3f), the loadings of this component show slopes between 1000 nm and 1300 nm and 2400 nm and 2500 nm. Peaks of 1730 nm and 1762 nm correspond to the lipids.

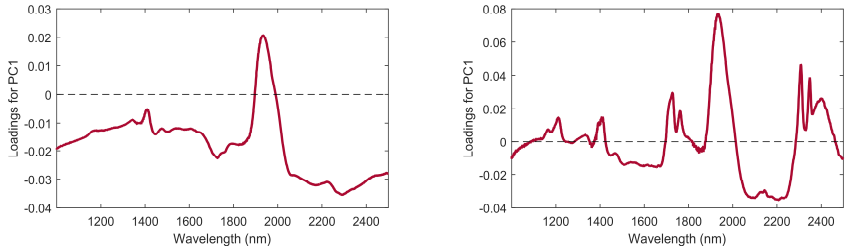
The description of the loadings of the first five components highlights that the variance of the repeated measures is related to physical phenomena. These physical phenomena induce essentially shifts in baselines on spectra. In the next components, chemical information appears. The repeatability error is therefore mainly of a physical nature, and to a lesser extent also due to a chemical difference, changing the diffusion of light in the medium.

4.4. Factor analysis after REP-ASCA (with k =5)

4.4.1. Species

Without or after orthogonal projection (Fig. 4a and 4b), the water content differentiates the two species by the peak of water located at 1930 nm. A second smaller peak located at 1398 nm corresponds to the hydrogen bond. It is also called weakly-hydrogen-bonded water [34]. Note that without orthogonal projection the loadings are all negative except the peak located at 1930 nm (Fig. 4a). After orthogonal projection according to the first 5 components of **W**, the loadings (Fig. 4b) are centered on zero. The orthogonal projection therefore removed the baseline.

Some information that did not appear before is now apparent (Fig. 4b). The three peaks at 2309 nm, 2348 nm and 2400 nm show a difference in fat



(a) (b)

Figure 4: Loadings on PC1 of the term $\mathbf{X}_{species}$: (a) Without pretreatment (b) by using the projection orthogonally to the 5 first components of $\mathbf{W}$.

content [7]. The peaks at 1728 nm and 1763 nm reveal a strong presence of fatty acids [28]. It's known that Arabica beans has higher lipid content than Robusta ones [28]. The two species can be differentiated by the lipid content. A less important peak at 1212 nm also shows a difference in caffeine and carbohydrates. A negative plateau between 2070 nm and 2238 nm shows a difference in proteins and chlorogenic acids content.

4.5. Impact of the number of spectra in $\mathbf{W}$

4.5.1. Number required for $\mathbf{W}$

In this study, $\mathbf{W}$ contains 24 spectra. This number is more than enough to describe the repeatability error. The impact of the number of observations on the structure of the subspace of $\mathbf{W}$ is studied using the RV coefficient between the matrices $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ with n ranging from 4 to 24 spectra. Bootstrap is performed to check that the results are not due to the random selection of observations available.

The graph 5 shows the evolution of the mean and standard deviation of this coefficient according to n. Mean values are higher than 0.90 for all the

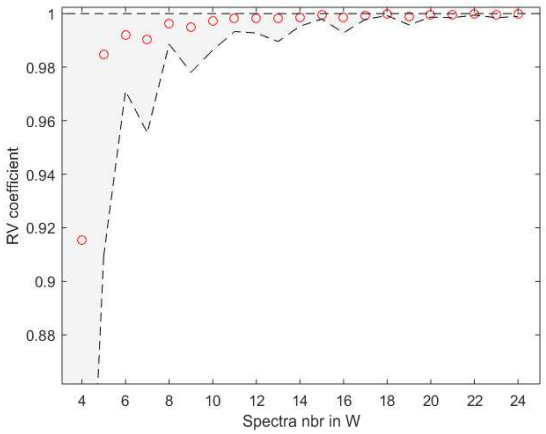


Figure 5: Evolution of the mean (red circle) and the standard deviation (grey area) of the RV coefficient between $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ according to the number n of spectra contained in $\mathbf{W}$

values of n. A plateau is visible with values close to 1 when n is greater than 10. The standard deviation displayed is about 0.18 when $n < 6$. When n increases ($n > 11$), standard deviation gradually decreases to values below 0.01. For a given n, a high value of the standard deviation means that the RV coefficient varies according to the n-th spectrum added in $\mathbf{W}$. The opposite is true, with a low standard deviation, the RV coefficient random varies slightly.

For $n < 8$, RV coefficients have low values meaning that the repeatability error is not correctly described. Additional observations must then be added. The increase of RV coefficient values with n is expected. Indeed, adding an observation has less impact in a dataset containing many observations. The plateau located from $n = 10$ shows that the structure of subspace $\mathbf{W}$ is stabilizing. From 10 observations, adding spectra in $\mathbf{W}$ does not provide

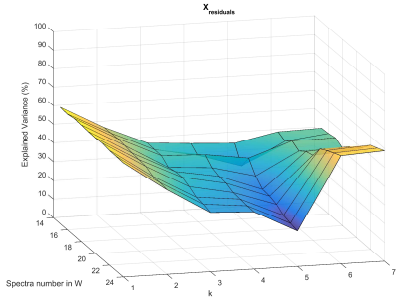


Figure 6: Explained variance of residuals according to number of projections and spectra number

more information on the repeatability error.

4.5.2. Impact on the selection of k

In this study, repeated measures are used to construct the $\mathbf{X}_s$ matrix. This matrix is used to describe the repeatability error. The impact of the number of spectra used for $\mathbf{X}_s$ matrix and the number of components used for orthogonal projection on the explained variance of residuals is illustrated in Fig. 6. The variance decreases up to the 5th component for a spectra number ranging from 15 to 24 in $\mathbf{X}_s$. The part of variance explained by the residuals is 37.31% and is 18.88% respectively for a number of 15 and 24 spectra in $\mathbf{X}_s$. The fewer spectra in $\mathbf{X}_s$, the greater residual variance. As a result, the repeatability error is then less described.

The p-values for each factor are presented in Fig. 7 according to the number of spectra used for $\mathbf{X}_s$ matrix and the number of components used for orthogonal projection. For the species factor (Fig. 7a), reducing the number of spectra from 24 spectra to 15 does not have an impact on the significance even after projection because species is the predominant factor

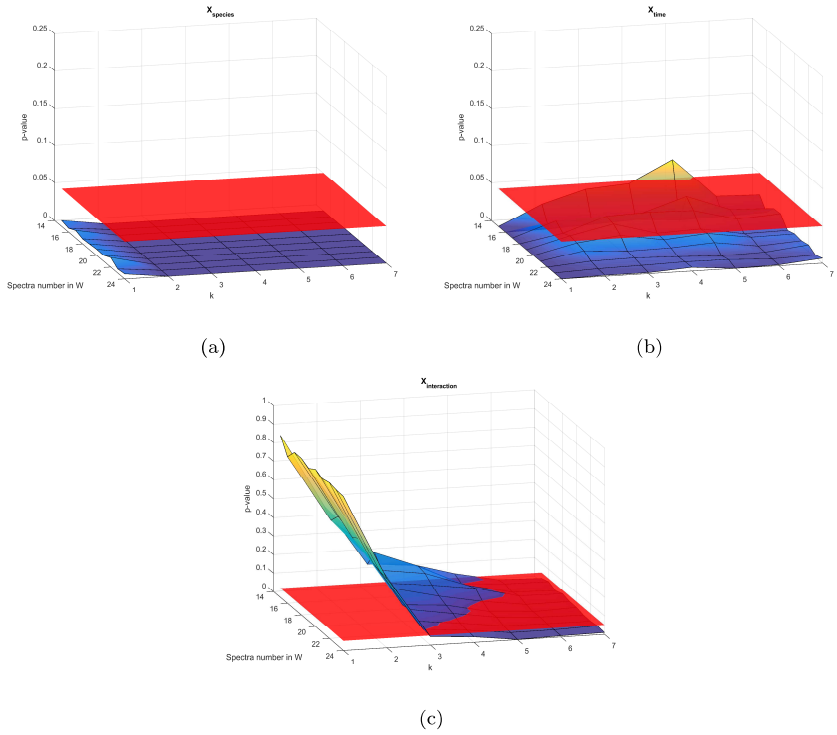


Figure 7: Impact of the spectra number in $\mathbf{X}_s$ and of the number k of components removed on p-values obtained for (a) Species, (b) Roasting time, (c) Interaction; The surface in red represents the value of 0.05 for the p-value.

in terms of portion of variance explained. For the roasting time (Fig. 7b) and the interaction terms (Fig. 7c), the number of spectra in $\mathbf{X}_s$ and the number of components used for orthogonal projection affect the results of the permutation tests. As presented in Fig. 7b, the roasting time term is not significant after a projection orthogonal to the first five components of $\mathbf{W}$ when $\mathbf{X}_s$ contains less than or equal to 15 spectra. For the interaction term (Fig. 7c), significance may change depending on the number of spectra used to build $\mathbf{X}_s$. Permutation tests may be strongly influenced when reducing the number of spectra in $\mathbf{X}_s$. Here, to study the interaction term and the roasting time term, it is necessary to have at least 20 spectra in $\mathbf{X}_s$. As a general, the number of observations must be sufficient to reduce the repeatability error. The more components to be removed, the higher the number of spectra required to be included in $\mathbf{X}_s$.

In the previous study [18], the Standard Normal Variate transform (SNV) [35] was used to reduce additive effects. This pretreatment is common in spectroscopy. It corrects the constant baselines over the whole dataset but SNV cannot correct non-constant (sloping or curving) baseline effects [36, 9].

The advantage of the proposed method is to reduce the repeatability error with more accuracy. Based on the analysis of repeated measures, the components of the repeatability error contain non-constant (sloping or curving) baselines. Indeed, they reflect diffusion and absorption phenomena of both physical and chemical nature.

5. Conclusion

In this study, a method to reduce the repeatability error (REP-ASCA) was tested on spectral data containing large repeatability error. Indeed, the interpretation of the factors through ASCA on such raw data (without correction) could mislead to wrong conclusions. The loadings of factor components contain then only the information related to the repeatability error. It was also illustrated how the significance of the factors can be influenced by the nature of the repeatability error.

The proposed method reduces these effects with the help of an additional set. The description of the repeatability error is then made through the study of repeated measurement introduced in the additional set. This description highlights the spectral regions related to the repeatability error. While repeatability error description is carried out in the column-space, reducing this error is realized in the row-space.

The choice of the number of components to be removed is done by studying the impact of projections orthogonally to these components on the various factors of the analysis of variance. Moreover, this choice can be improved by the inspection of the loadings of the main components of the repeatability error. While removing the error due to repeatability, the results obtained with REP-ASCA allow to highlight the spectral regions due to the factors of interest.

REP-ASCA was tested on a dataset containing a repeatability error mainly caused by samples with highly variable physical specificities and associated measurement errors (particle size, angle of view, etc). It would be interesting to apply REP-ASCA to another dataset containing other kinds of

repeatability errors. In addition to reducing repeatability error, REP-ASCA could remove other undesired effects (temporal effect, specific chemical content) by incorporating other components related to these effects. Futhermore, the REP-ASCA methodology could correct any type of unwanted effects, provided that the spectral subspace generated by these effects can be identified. A lot of methods can be used to carry out this identification [37].

On the practical side, adding an additional set to reduce repeatability error may considerably reduce the number of repetitions per modality in the set for the analysis of variance. Using this method, acquisition protocols should include observations to describe these errors independently of the observations for analysis of variance. This is required to make the measures for the analysis of variance more robust. However, a sufficient number of observations should be kept to study the significance of the factors.

References

[1] F. E. Grubbs, On estimating precision of measuring instruments and product variability, Journal of the American Statistical Association 43 (1948) 243–264.

[2] J. N. Miller, J. C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearson Education, 2005. Google-Books-ID: Efx77dxOC3sC.

[3] R. A. Fisher, The design of experiments, Oliver And Boyd; Edinburgh; London, 1937.

[4] G. W. Oehlert, A first course in design and analysis of experiments, W.H. Freeman, New York, 2000.

[5] R. A. Fisher, Statistical methods for research workers, Genesis Publishing Pvt Ltd, 2006.

[6] H. Scheffé, The analysis of variance, The analysis of variance, Wiley, Oxford, England, 1959.

[7] P. Williams, K. Norris, Near-infrared technology in the agricultural and food industries., American Association of Cereal Chemists, Inc., 1987.

[8] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, L. M. Buydens, Breaking with trends in pre-processing?, TrAC Trends in Analytical Chemistry 50 (2013) 96–106.

[9] Å. Rinnan, F. v. d. Berg, S. B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry 28 (2009) 1201–1222.

[10] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, J.-S. S. White, Generalized linear mixed models: a practical guide for ecology and evolution, Trends in Ecology & Evolution 24 (2009) 127–135.

[11] S. L. Zeger, K.-Y. Liang, Longitudinal data analysis for discrete and continuous outcomes, Biometrics (1986) 121–130.

[12] M. Anderson, C. T. Braak, Permutation tests for multi-factorial analysis of variance, Journal of statistical computation and simulation 73 (2003) 85–113.

[13] R. G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Analytical and Bioanalytical Chemistry 409 (2017) 5891–5899.

[14] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048.

[15] R. Climaco-Pinto, A. Barros, N. Locquet, L. Schmidtke, D. Rutledge, Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability, Analytica Chimica Acta 653 (2009) 131–142.

[16] R. G. Shaw, T. Mitchell-Olds, Anova for Unbalanced Data: An Overview, Ecology 74 (1993) 1638–1645.

[17] A. K. Smilde, H. C. Hoefsloot, J. A. Westerhuis, The geometry of ASCA, Journal of Chemometrics 22 (2008) 464–471.

[18] S. De Luca, M. De Filippis, R. Bucci, A. D. Magrì, A. L. Magrì, F. Marini, Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics, Microchemical Journal 129 (2016) 348–361.

[19] D. J. Vis, J. A. Westerhuis, A. K. Smilde, J. van der Greef, Statistical validation of megavariate effects in ASCA, BMC Bioinformatics 8 (2007) 322.

[20] M. J. Anderson, Permutation tests for univariate or multivariate analysis of variance and regression, Canadian Journal of Fisheries and Aquatic Sciences 58 (2001) 626–639.

[21] H. A. Kiers, SCA: A Program for Simultaneous Components Analysis of Variables Measured in Two Or More Populations: user's Manual, iec ProGamma, 1990.

[22] K. Van Deun, A. K. Smilde, M. J. van der Werf, H. A. Kiers, I. Van Mechelen, A structured overview of simultaneous component based data integration, Bmc Bioinformatics 10 (2009) 246.

[23] F. Marini, D. de Beer, E. Joubert, B. Walczak, Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach, Journal of Chromatography A 1405 (2015) 94–102.

[24] ISO-5725-2 1994(en), International Organization for Standardization, 1994, Accuracy (trueness and precision) of measurement methods and results— Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Standard, ????

[25] Y. Escoufier, Le Traitement des Variables Vectorielles, Biometrics 29 (1973) 751.

[26] H. Abdi, RV coefficient and congruence coefficient, Encyclopedia of measurement and statistics 849 (2007) 853.

[27] L. Alessandrini, S. Romani, G. Pinnavaia, M. D. Rosa, Near infrared spectroscopy: An analytical tool to predict coffee roasting degree, Analytica Chimica Acta 625 (2008) 95–102.

[28] I. Esteban-Díez, J. González-Sáiz, C. Pizarro, Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy, Analytica Chimica Acta 525 (2004) 171–182.

[29] J. Ribeiro, M. Ferreira, T. Salva, Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy, Talanta 83 (2011) 1352–1358.

[30] K. A. Bakeev (Ed.), Process analytical technology: spectroscopic tools and implemented strategies for the chemical and pharmaceutical industries, Wiley, Chichester, West Sussex, 2nd ed edition, 2010. OCLC: ocn473478595.

[31] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbeke, C. K. Mann, Handbook of Chemometrics and Qualimetrics: Part A, Applied Spectroscopy 52 (1998) 302A.

[32] J. Workman, A. W. Springsteen (Eds.), Applied spectroscopy: a compact reference for practitioners, Academic Press, San Diego, 1998.

[33] W. A. Luck, Structure of water and aqueous solutions, Verlag Chemie, 1974.

[34] A. A. Gowen, R. Tsenkova, C. Esquerre, G. Downey, C. P. O'Donnell, Use of near Infrared Hyperspectral Imaging to Identify Water Matrix Co-Ordinates in Mushrooms (*Agaricus Bisporus*) Subjected to Mechanical Vibration, Journal of Near Infrared Spectroscopy 17 (2009) 363–371.

[35] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy 43 (1989) 772–777.

[36] R. P. Cogdill, C. R. Hurburgh, G. R. Rippke, S. J. Bajic, R. W. Jones, J. F. McClelland, T. C. Jensen, J. Liu, others, Single-kernel maize analysis by near-infrared hyperspectral imaging, Transactions of the ASAE 47 (2004) 311.

[37] J.-M. Roger, J.-C. Boulet, A review of orthogonal projections for calibration, Journal of Chemometrics (2018) e3045.

Communications

B.1 Communications aux congrès

- EFITA (2019) : Removing spatial effects of spectral dataset acquired into an experimental design by using multivariate analysis of variance.
- Chimiometrie (2019) : Predicting maize yield of new varieties from temporal-spectral data using multibloc-analysis.
- SFPT (2018) : Multivariate analysis of variance of vegetation spectra dataset included into an experimental design by using ANOVA-SCA and ANOVA-Target Projection.
- EFITA (2017) : The impact of the spatial resolution of highly resolved spectral data on pansharpening methods to reconstruct a hyperspectral image.
- HelioSpir (2017) : ANOVA-Simultaneous component analysis on vegetation spectra data acquired into an experimental design.

B.2 Articles publiés dans des revues scientifiques à comité de lecture

- (*Analytica Chimica Acta* - 2019 - Accepted) : Reduction of repeatability error for Analysis of variance-Simultaneous Component Analysis (REP-ASCA) : Application to NIR spectroscopy on coffee sample - Maxime Ryckewaert, Nathalie Gorretta, Fabienne Henriot, Federico Marini, Jean-Michel Roger
- (*Applied Optics* - 2019 - Accepted) : Combining light polarization and speckle measurements with multivariate analysis to predict bulk optical properties of turbid media - Daphné Héran, Maxime Ryckewaert, Yannick Abautret, Myriam Zerrad, Claude Amra, Ryad Bendoula
- (*Sensors* - 2019 - Accepted) : Coupling waveguide-based micro-sensors and spectral multivariate analysis to improve spray deposit characterization in agriculture - Anis Taleb Bendiab, Maxime Ryckewaert, Daphné Heran, Raphael Escalier, Raphael Kribich, Caroline Vigreux, Ryad Bendoula

Table des figures

1.1 Exemples de micro-parcelles. 14
1.2 Fécondation croisée entre parents mâle et femelle. 16
1.3 Schéma général de sélection du maïs. 17

2.1 Représentation couleur de l'image hyperspectrale de référence : betterave sucrière. 36
2.2 Méthodologie pour évaluer l'influence de la résolution spatiale de l'image spectrale sur les qualités de reconstruction des méthodes de pan-sharpening. 37
2.3 Méthodologie pour évaluer l'influence de la résolution spatiale de l'image panchromatique sur les qualités de reconstruction des méthodes de pan-sharpening. 38
2.4 Evolution des indices pour les différentes méthodes de pan-sharpening testées en fonction de la résolution spatiale N : (a) CC, (b) SAM, (c), RMSE. 39
2.5 Qualité de reconstruction pixel à pixel selon l'indice SAM pour différentes méthodes de pan-sharpening : (a) BT, (b) SFIM, (c), CNMF. 41
2.6 Le coefficient CC par longueur d'onde. 42
2.7 Comparaison des durées pour reconstruire une image en fonction de N pour : (a) l'ensemble des méthodes sauf HySure (b) la méthode HySure. 43
2.8 Comparaison des méthodes de pan-sharpening en fonction de M suivant les indices : (a) CC, (b) RMSE, (c), SAM (deg). 44
2.9 Comparaison de la durée de calculs des méthodes de pan-sharpening en fonction de la résolution spatiale M de l'image panchromatique avec : (a) une échelle adaptée à la majorité des méthodes, (b) selon une échelle adaptée à HySure. 45
2.10 Evolution des indices : (a) CC, (b) RMSE, (c), SAM (deg) en fonction du taux de recouvrement pour les approches CNMF-KNN (rouge) et CNMF-PLS (bleu). 49
2.11 Indices SAM pixel à pixel pour les deux méthodes proposées : (a) CNMF-KNN, (b) CNMF-PLS. 50

2.12 Évolution de l'indice CC en fonction de la longueur d'onde pour les deux approches proposées. 50
2.13 Indice SAM pour chaque pixel selon l'utilisation les méthodes : (a) KNN, (b) PLS. 51
2.14 Indice CC par bandes spectrales des méthodes proposées après la fusion entre l'image hyperspectrale. 52
2.15 Synthèse du chapitre 2 : Méthode proposée pour reconstruire une image hyperspectrale à partir d'une image panchromatique ou multispectrale et des informations spectrales sur cette image. 55

3.1 Représentation matricelle d'une collection d'observations. 63
3.2 Représentation des espaces : (a) formé par les observations (ou column-space) (b) formé par les variables (ou row-space). 64
3.3 Spectres simulés : (a) avec le facteur A significatif (b) masquant l'effet du facteur A 69
3.4 Évolution de la part de variance du facteur A en fonction du nombre de composantes retirées. 70
3.5 Loadings obtenus sur la première composante du facteur A après (1) - courbe rouge, ASCA sur les données de références ; (2) - courbe bleu, ASCA sur les données contenant l'erreur de répétabilité, courbe en bleu ; (3) - courbe jaune, REP-ASCA sur les données contenant l'erreur de répétabilité 70
3.6 Loadings des composantes de l'erreur de répétabilité : (a) simulés (b) retrouvés avec REP-ASCA 71
3.7 Synthèse du chapitre 3 : L'illustration de la méthode REP-ASCA. 73

4.1 Dispositif de mesure des spectres. 80
4.2 Support de mesure. 80
4.3 Points de mesures établis sur une micro-parcelle semée en 4 rangs (rouge). 81
4.4 Spectres de réflectance des mesures spectrales effectuées à la date 3 de l'année 2018. 87
4.5 Impact des projections sur les variances expliquées des facteurs de l'analyse de variance. 87
4.6 Description de l'erreur de répétabilité : (a) loadings de la première composante de $\mathbf{W}$, (b) loadings de la deuxième composante de $\mathbf{W}$. 89
4.7 Description du terme traitement : (a) Loadings de la composante principale, (b) Scores obtenus par génotype sur la composante principale. . . . 90

4.8 Description du terme génotype : (a) Loadings de la première composante principale, (b) Loadings de la deuxième composante principale. 91
4.9 Scores des génotypes sur les deux premières composantes (PC1 et PC2) du terme génotype. 92
4.10 Description du terme d'interaction :(a) Loadings de la première composante principale, (b) Loadings de la deuxième composante principale. . . 93
4.11 Scores obtenus sur les deux premières composantes PC1, PC2 du terme d'interaction. 94
4.12 Scores-plot et ellipses de confiance associées sur les deux premières composantes des génotypes :(a) E et G, (b) E et F. 95
4.13 Transformation des scores en coordonnées polaires avec rhô et thêta. . . 98
4.14 Scores en fonction du pourcentage de pertes (Loss). 100

5.1 Evolution des scores moyens sur la composante du terme traitement en fonction des dates d'acquisition. 110
5.2 Évolution des scores obtenus sur les deux premières composantes du terme génotype aux différentes dates pour les génotypes A, E, G et H. La numérotation indique les dates d'acquisition. 111
5.3 Évolution des scores obtenus sur la première composante du terme interaction en fonction des dates d'acquisition et pour l'ensemble des génotypes. 113
5.4 Evolution des moyennes et des écarts-types des scores par traitement en fonction de la date. 117
5.5 Evolution des scores obtenus sur les deux premières composantes selon les dates pour un autre environnement. 118
5.6 Evolution des scores obtenus sur la première composante du terme d'interaction selon les dates pour un autre environnement 119
5.7 Scores par génotype obtenus sur l'environnement de 2017 en fonction du pourcentage de pertes en rendement de l'environnement 2018. 120
5.8 Synthèse du chapitre 5 : Application des proxis obtenus par REP-ASCA sur d'autres dates et sur un autre environnement. 122

A.1 Plan de vol du drone embarquant AirStrip pour la campagne de 2019 . . 130
A.2 Première composante du terme d'interaction issue de la campagne 2018 adaptée au spectromètre Ocean Optics 131
A.3 Scores obtenus pour la campagne 2019 par génotypes 133
A.4 Illustration des scores obtenus sur les micro-parcelles correspondant aux génotypes (a) I et (b) E 133

Liste des tableaux

2.1 Spécifités des paramètres g_k et ω_i pour les méthodes de pan-sharpening appartenant à la classe CS. 30

2.2 Spécifités des paramètres des méthodes de pan-sharpening appartenant à la catégorie MRA. 31

2.3 Valeurs des indices CC, SAM et RMSE pour les deux approches testées en utilisante une image multispectrale. 51

4.1 Moyenne, écart type et héritabilité des variables rendement et humidité obtenues à la récolte pour les deux conditions d'irrigation. 82

4.2 ANOVA sur la variable rendement. 83

4.3 Rendement moyen par traitement et pourcentage de pertes pour chaque génotype. 84

4.4 Déficit de pression de vapeur (VPD) et tensiométrie (cbar) des dates d'acquisitions de l'année 2018. 86

4.5 Températures (°C) maximales de la journée et pour les 3 jours précédents des dates d'acquisitions de l'année 2018. 86

4.6 Corrélation de la variable rendement avec les scores bruts du facteur traitement scores $\mathbf{S}$ et une différence de scores par génotype entre les deux traitements $\mathbf{S}_{diff}$. 97

4.7 Corrélation des variables agronomiques avec les scores obtenus pour le terme génotype. 98

4.8 Corrélation des variables agronomiques avec les scores obtenus pour le terme d'interaction génotype-traitement. 99

5.1 Corrélations entre l'indicateur thêta et le pourcentage de perte en rendement. 112

5.2 Coefficients de corrélation entre les scores par date (obtenus avec la première composante de la date 3) et le pourcentage de perte en rendement. 114

5.3 Déficit de pression de vapeur (VPD), tensiométrie (cbar) pour chaque traitement et différence de tensions entre les traitements pour l'ensemble des dates d'acquisitions de l'année 2017. 115
5.4 Valeurs de rendements moyens par traitement et pourcentage de pertes (Loss) pour chacun des génotypes. 116

A.1 Déficit de pression de vapeur (VPD), tensiométrie (cbar) pour la campagne de l'année 2019. 131
A.2 Valeurs de rendements moyens par traitement et pourcentage de pertes (Loss) pour chacun des génotypes. 132