ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

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

PAR
CLÉMENT CHION

PROGRAMMATION GÉNÉTIQUE APPLIQUÉE À L'IMAGERIE HYPERSPECTRALE POUR L'ÉVALUATION D'UNE VARIABLE BIOPHYSIQUE AU SEIN D'UNE GRANDE CULTURE : CAS DE L'AZOTE DANS UN CHAMP DE MAÏS

MONTRÉAL, LE 20 OCTOBRE 2005

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Jacques-André Landry, directeur de mémoire
Département de génie de la production automatisée à l'École de Technologie Supérieure

M. Mohamed Cheriet, président du jury
Département de génie de la production automatisée à l'École de Technologie Supérieure

M. Tony Wong, examinateur
Département de génie de la production automatisée à l'École de Technologie Supérieure

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

LE 6 DÉCEMBRE 2005

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

RÉSUMÉ DE 150 MOTS

PROGRAMMATION GÉNÉTIQUE APPLIQUÉE À L'IMAGERIE HYPERSPECTRALE POUR L'ÉVALUATION D'UNE VARIABLE BIOPHYSIQUE AU SEIN D'UNE GRANDE CULTURE : CAS DE L'AZOTE DANS UN CHAMP DE MAÏS

Clément CHION

L'imagerie hyperspectrale de télédétection offre d'innombrables opportunités pour la gestion durable des ressources naturelles. L'agriculture de précision est une approche récente qui prend en considération l'hétérogénéité biophysique des cultures, lors de l'application d'intrants (engrais, herbicides...). Nous proposons une nouvelle méthode fondée sur les principes de la programmation génétique et des indices de végétation; l'objectif est d'élaborer un modèle décrivant une variable biophysique d'un champ, pour évaluer précisément sa variabilité et agir localement. La validation de notre approche est réalisée sur des mesures d'azote (variable biophysique étudiée) relevées dans un champ-test de maïs de l'Université McGill (Montréal). Le meilleur modèle obtenu explique 84.83% de la variance d'un jeu de données non apprises avec une erreur de généralisation de 14.34%, améliorant ainsi les résultats de la littérature. Un autre résultat important est que les performances des modèles obtenus dépendent moins de la taille du jeu de données d'apprentissage que de leur précision.

PROGRAMMATION GÉNÉTIQUE APPLIQUÉE À L'IMAGERIE HYPERSPECTRALE POUR L'ÉVALUATION D'UNE VARIABLE BIOPHYSIQUE AU SEIN D'UNE GRANDE CULTURE : CAS DE L'AZOTE DANS UN CHAMP DE MAÏS

Clément Chion

SOMMAIRE

Un enjeu majeur de la télédétection est l'extraction d'information pertinente contenue dans les données. Le récent développement de l'imagerie hyperspectrale a fait exploser le volume de données à explorer et par conséquent, de nouvelles techniques d'analyse sont requises. En agriculture de précision, l'avènement et la démocratisation des technologies d'acquisition de mesures spectrales multibandes laisse entrevoir de grands espoirs dans l'optique d'une gestion plus raisonnée des ressources. En effet, les propriétés spectrales des plantes et de leurs composants étant largement étudiées, l'extension des connaissances de l'échelle de la plante à celle de la canopée semble prometteuse. Toutefois, des facteurs tels que l'irradiance, l'humidité de l'air ou encore l' « impureté » des pixels compliquent le transfert des connaissances lors du changement d'échelle. Pour remédier à ces problèmes, des indices de végétation (IV), définis par des combinaisons arithmétiques de bandes spectrales, ont été développés afin d'isoler des variables biophysiques de la canopée. Dans notre étude, nous cherchons un IV corrélé à la variabilité en azote d'un champ de maïs, au moyen d'un algorithme basé sur la programmation génétique entraîné sur des mesures prises en champ. Par cette technique, nous avons déterminé un modèle qui prédit la quantité d'azote dans le champ avec $R^2 = 84.83\%$ et ce avec une erreur relative RMSE = 14.34%. Ce résultat obtenu sur nos données est meilleur que ceux des indices et modèles de la littérature; la meilleure performance étant celle d'un modèle de Hansen et al. prédisant l'azote avec $R^2 = 70.23\%$ et une erreur RMSE = 18.03%. L'analyse d'un potentiel transfert technologique conduit à un autre résultat majeur : la précision du modèle descriptif de la variable biophysique dépend moins de la taille de l'ensemble de données d'apprentissage que de la précision de celles-ci. Il semble que l'on ne soit pas encore en mesure de trouver des modèles généraux indépendants des facteurs externes. En attendant, à l'aide d'algorithmes de recherche, l'investigation de données hyperspectrales couplées à des mesures in situ permet de contourner ce problème en gardant un niveau de performance élevé.

GENETIC PROGRAMMING APPLIED TO HYPERSPECTRAL IMAGERY FOR BIOPHYSICAL VARIABLE ASSESSMENT WITHIN A LARGE SCALE CULTURE: CASE OF NITROGEN WITHIN A CORNFIELD

Clément Chion

ABSTRACT

One of the main issues of remote sensing is the extraction of relevant information from a data set. Recent development of hyperspectral tools has considerably increased the amount of available data and consequently, new techniques for data mining are required. In precision farming, emergence and democratization of hyperspectral imagery gives rise to great hopes by providing powerful tools to set up more reasonable management. Indeed, spectral properties of plants and their components being well studied, extrapolation of this knowledge from plant to canopy scale appears to be promising. However, many external factors like air humidity, irradiance or effect of pixel resolution bring some noise and make information extraction more complex at canopy scale. An answer to this problem can be brought by vegetation indices (IV), defined as simple arithmetic combinations of spectral bands. One of the goals of these IV is to bring out a specific canopy biophysical parameter. In our study, we try to find an IV correlated with nitrogen variability through a cornfield canopy, by means of a genetic programming-based algorithm, trained with in situ measures. This approach led us to find a model predicting nitrogen levels through the field with a coefficient of determination $R^2 = 84.83\%$ and a relative error RMSE = 14.34%. This result obtained with our data set improves all others models found in articles; the best of them given by Hansen & al. predicting nitrogen with $R^2 = 70.23\%$ and RMSE = 18.03%. The other important result is that model precision less depends on dataset size than on training data accuracy. At present, it doesn't yet seem possible to find a general model for nitrogen assessment, efficient in all of real situations. Meanwhile, coupling "ground truth" with hyperspectral data can lead to great levels of efficiency when investigations are made with specific search algorithms.

REMERCIEMENTS

Tout d'abord, je souhaite remercier mon directeur M. Jacques-André Landry qui a été un soutien sur plusieurs plans et qui, en m'offrant sa confiance, m'a donné la possibilité de changer d'orientation en douceur.

Je souhaite bien évidemment souligner l'importance de la bonne humeur des Liviens dans le bon déroulement de ces deux années. Le LIVIA m'a procuré bien plus qu'une atmosphère de travail.

Je dédie ce travail à toute ma famille qui est mon soutien moral de tous les instants ainsi qu'à mon père qui je l'espère apprécie les chemins que j'emprunte.

Enfin, je remercie ma grenouille qui a le mérite de me supporter.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES GRAPHIQUES

Page

LISTE DES FIGURES

Page

LISTE DES SIGLES ET ABRÉVIATIONS

ADN	Acide Désoxyribonucléique
ARVI	Atmospherically Resistant Vegetation Index
CARI	Chlorophyll Absorption Ratio Index
CASI	Compact Airborne Spectrometer Imager
CCD	Charge-Coupled Device
Chl_{conc}	Concentration en chlorophylle en mg par g de feuilles fraîches
$Chl_{densité}$	Densité de chlorophylle en mg par m² de sol
CRESS	Centre for Research in Earth and Space Science
CV(X)	Coefficient de Variation d'une série d'observations de la variable X. $CV(X) = \frac{\text{écart - type}(X)}{\text{moyenne}(X)} = \frac{\sigma(X)}{\mu(X)}$
EQMN	Erreur Quadratique Moyenne Normalisée (⇔ NMSE)
GBM	Green biomass, en grammes de masse fraîche de végétation par m² de sol
GDVI	Green Difference Vegetation Index
GLAI	Green Leaf Area Index, m²/m²
GPS	Global Positioning System
IFC	Intensive Field Campaign
IFOV	Instantaneous Field Of View
ISF	Indice de Surface Foliaire en m² de feuilles vertes par m² de sol (=LAI)
IV	Indice de Végétation
KNN	K-Nearest-Neighbors (= K-plus-proches-voisins, KPPV)
LAD	Leaf Angle Distribution
LAI	Leaf Area Index, m²/m²
MCARI	Modified Chlorophyll Absorption Ratio Index
MSR	Modified Simple Ratio
N_{conc}	Concentration d'azote en mg par g de feuille sèche

$N_{densité}$	Densité d'azote dans les feuilles en g par m² de sol
NDVI	Normal Difference Vegetation Index
NIR	Near Infrared
NMSE	Normalized Mean Square Error (erreur quadratique moyenne normalisée)
PIR	Proche Infrarouge
PWC	Plant Water Content
R^2	Coefficient de détermination entre deux séries numériques
RDVI	Renormalized Difference Vegetation Index
RMSE	Root Mean Square Error (racine carrée de l'erreur quadratique moyenne)
$RMSE_{abs}$	Absolute Root Mean Square Error
$RMSE_{\%}$	Relative Root Mean Square Error
SARVI	Soil and Atmospherically Resistant Vegetation Index
SAVI	Soil-Adjusted Vegetation Index
SIG	Système d'Information Géographique
SPAD	Specialty Products Agricultural Division
TVI	Triangular Vegetation Index
VRT	Variable-Rate Technologies
WI	Water Index

CHAPITRE 1

MISE EN CONTEXTE

1.1 Contexte

1.1.1 Les besoins des plantes

À l'état naturel ou bien au sein de cultures agricoles, les plantes sont en perpétuelle compétition pour ce qui est de l'accès aux ressources vitales dont les principales sont a) l'eau; b) la lumière et c) les nutriments tels que le phosphore et l'azote en particulier.

Dans certaines cultures, l'apport des eaux de pluie et des nappes d'eau souterraines est complété par l'irrigation artificielle pour combler le besoin des plants vis-à-vis de cette ressource. L'espace cultivable disponible restreint de façon intrinsèque le nombre de plants qui pourront avoir accès à la lumière. Enfin, les nutriments présents dans le sol constituent les facteurs limitants sur lesquels l'agriculteur va agir pour combler les carences. Ainsi, à plusieurs reprises au cours de l'année, l'exploitant vient ajouter des engrais chimiques (azote principalement) pour répondre aux besoins des plants et leur permettre ainsi un meilleur développement. L'azote est l'élément nutritif essentiel des plantes qu'elles extraient du sol sous forme de nitrate (NO_3). Cet élément sert à la synthèse de la plupart des protéines, des molécules d'ADN, des vitamines, des hormones, des enzymes et également de la chlorophylle dont les plantes sont constituées majoritairement. Une carence en azote entraîne généralement une décoloration des feuilles et une croissance végétale faible au printemps. À l'inverse, en excès il conduit à une croissance exagérée du feuillage aux dépens de la fructification et provoque un manque de rigidité des plants. Ces constats soulèvent l'importance de fournir la dose d'azote adéquate pour permettre à la végétation de se développer de façon optimale.

1.1.2 État des sols

La dégradation des sols arables est un problème réel au Canada ainsi que dans de nombreux pays industrialisés. Plusieurs rapports ont été écrits pour sensibiliser les producteurs agricoles canadiens. Un important message d'alarme sur ce sujet fut le rapport sénatorial de 1984 intitulé "Nos sols dégradés: le Canada compromet son avenir" écrit par l'Honorable Herbert O. Sparrow [1]. Dans ce document, le sénateur Sparrow avait déclaré qu'à moins de prendre rapidement des mesures, le Canada allait perdre une grande partie de ses terres agricoles au cours des cent prochaines années. Depuis, des dispositions ont été prises pour aller à l'encontre de ce phénomène de détérioration mais il reste encore beaucoup de chemin à parcourir pour aller vers une agriculture dite durable ou responsable. Si elle se pratique de façon classique, l'agriculture n'a pas d'avenir; il ne sera pas possible de produire davantage chaque année comme cela se fait depuis 20 ans. Des politiques de développement durable sont mises en place pour apporter des réponses à ce problème.

L'agriculture responsable a pour but de répondre aux besoins actuels sans compromettre la possibilité, pour les générations à venir, de subvenir aux leurs. Pour atteindre cet objectif, il faut agir sur les facteurs qui détériorent l'état des sols. Ces facteurs sont nombreux et les principaux sont l'érosion éolienne et l'érosion hydrique des sols, leur perte en matière organique, leur changement de structure, l'augmentation de leur salinité et la contamination agrochimique, notamment celle due aux excédents d'engrais azotés.

Les terres agricoles peuvent être contaminées de plusieurs façons, les principales étant les dépôts atmosphériques de déchets industriels et l'application directe de produits chimiques en excès. La motivation de ce projet est incidemment ce qui est appelé l'agriculture de précision. La philosophie de cette approche est d'effectuer des traitements (chimiques ou naturels) ciblés et variables sur les champs et non pas globaux comme ce qui se fait en agriculture classique; **l'objectif étant d'appliquer la bonne**

dose, au bon endroit et au bon moment. Un champ est une entité hétérogène et de ce fait, la présence des nutriments dans le sol ne peut être uniforme sur une zone de plusieurs centaines d'hectares (la taille moyenne d'une exploitation agricole au Canada était d'environ 400 hectares en 1996 d'après la Fédération Canadienne de l'Agriculture[1]). Pourtant, jusqu'à présent, les champs de grandes cultures sont majoritairement traités en excès. Les produits chimiques excédentaires (inutilisés par la végétation), peuvent soit demeurer dans le sol, soit aller dans les eaux de surface (par ruissellement) ou encore dans les eaux souterraines (par lessivage). Cela a pour conséquence de contaminer le sol, les eaux de surface ou les eaux souterraines et donc de compromettre les récoltes futures ainsi que l'utilisation de ces ressources pour tout autre usage.

1.1.3 Agriculture de précision

Dans le passé, la taille des champs était petite et les parcelles étaient sélectionnées pour des applications spécifiques en fonction des caractéristiques du sol. Généralement, les terres les moins riches étaient converties en prairies tandis que les plus productives servaient à la culture de céréales.

La mécanisation des procédés agricoles a permis une augmentation de la taille des exploitations induisant une variabilité intraparcellaire des caractéristiques intrinsèques (composition chimique du sol, élévation du terrain, rendement...). L'agriculture de précision prend en compte cette variabilité; elle permet donc un retour aux techniques d'autrefois contrairement à l'approche traditionnelle qui traite l'ensemble de la parcelle uniformément. Le développement de cette « nouvelle » façon de pratiquer l'agriculture est relié à la démocratisation des nouvelles technologies comme le Système de Positionnement par Satellite (Global Positioning System, GPS), les capteurs embarqués sur les machines agricoles, les technologies VRT (« Variable Rate Technology »)

[1] http://www.cfa_fca.ca/francais/lagriculture_au_canada/structure_et_aspects_financiers.html

définies comme les machines et systèmes permettant d'appliquer des intrants (pesticides, herbicides, graines, fertilisants) de façon variable, la télédétection ou encore les Systèmes d'Information Géographique (SIG) et la Géomatique. Les données collectées (au sol et par télédétection) sont analysées grâce aux SIG; des cartes de prescription sont alors créées. Les ordinateurs embarqués à bord des machines agricoles couplés aux technologies VRT et aux GPS qui les équipent permettent d'élaborer une stratégie d'application variable d'intrants, comme montré à la figure 1.

En ciblant précisément les besoins spécifiques des plants, il est possible de mieux y répondre; par conséquent, cela permet la diminution des quantités de produits chimiques résiduels ce qui réduit incidemment la pollution des sols et de l'eau souterraine. Outre le respect environnemental, un autre bénéfice majeur de l'agriculture de précision est la diminution des coûts pour l'exploitant par la réduction de la quantité d'intrants; de plus, une meilleure connaissance des caractéristiques du champ peut lui permettre de mieux gérer les risques liés à ses choix. Le rendement en agriculture est défini comme le rapport entre la production végétale obtenue (en poids, volume ou en nombre d'individus) et une unité de surface déterminée. Par exemple, si l'agriculteur est capable de prévoir les rendements des différentes parcelles (zones) de son exploitation (par la connaissance de l'historique des récoltes par zone), il va pouvoir établir un ratio moyen *rendement / hectare* et ainsi procéder à des choix d'investissement différents pour ses parcelles situées au-dessus et celles situées en dessous de cette valeur de référence [2]. Les bons résultats de l'agriculture de précision et l'amélioration continuelle des technologies connexes en font un domaine en croissance.

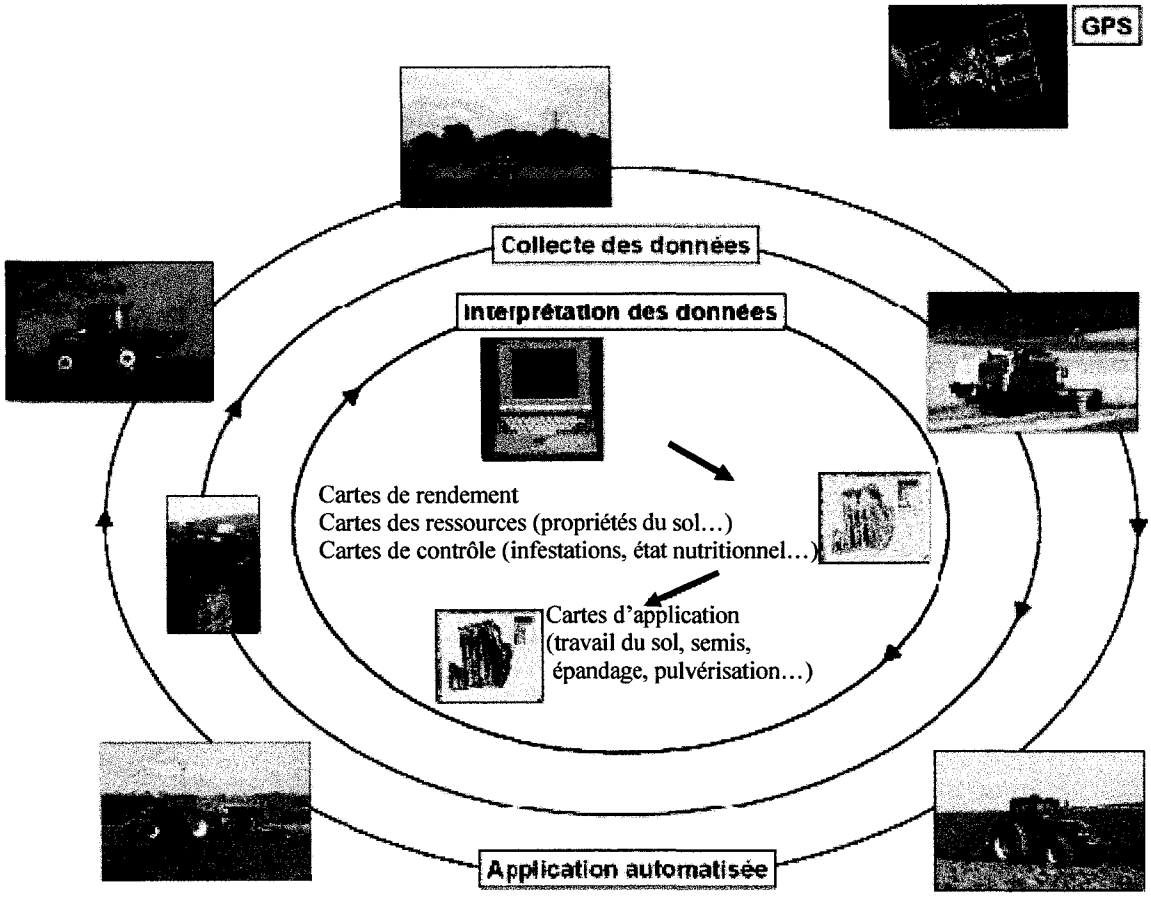


Figure 1 Étapes de l'agriculture de précision (Faculté des sciences agronomiques de Gembloux)

Un des enjeux de l'agriculture de précision est d'évaluer, le plus justement possible, le besoin des plants en nutriments, à divers moments de la saison et ce sur l'ensemble d'une parcelle. Plusieurs solutions existent pour effectuer le zonage des terres selon leurs caractéristiques, les plus répandues étant l'utilisation de capteurs embarqués et celle de la télédétection (satellite ou bien aéroportée). Les technologies VRT requièrent une information en entrée et celle-ci est fournie soit en temps réel par capteur soit par une carte établie par analyse préliminaire, notamment de données de télédétection. De nombreux capteurs existent, les plus communs étant les capteurs de rendement (activés au moment de la récolte), de profondeur (utiles pour connaître le volume d'eau

disponible pour la culture) et de texture. D'autres capteurs sont à l'étude dans les centres de recherche.

La télédétection aérienne est en train de devenir un outil majeur en agriculture de précision pour trois raisons principales qui sont :

a. procédé non destructif;
b. procédé permettant d'obtenir de l'information sur tout l'ensemble du champ;
c. acquisition rapide.

Toutefois, elle se heurte encore à deux problèmes majeurs qui sont le prix élevé des images à l'achat et le manque de connaissance pour l'extraction d'information précise de ces données. Les détails sur la place de la télédétection en agriculture de précision seront présentés dans le chapitre consacré à la revue de littérature. Pour plus d'information sur l'agriculture de précision, le lecteur pourra se référer au site placé en note de pied de page[2].

1.2 But du projet

Dans le cadre de ce projet, nous nous intéressons à trouver une « autre méthode » d'analyse du contenu d'une image hyperspectrale, avec pour objectif d'évaluer la variabilité en azote au sein des plants de maïs d'un champ. Les données disponibles sont les suivantes :

a. une image numérique hyperspectrale de télédétection (cf. ANNEXE 1) d'un champ de maïs, acquise par le capteur aéroporté CASI (Compact Airborne Spectrometer Imager);

[2] http://precision.agri.umn.edu/links.shtml

b. des mesures de réflectance prises au sol sur des plants (capteur ASD et capteur SPAD 502 de Minolta décrites à la section 3.3.1), peu de temps après l'acquisition par CASI. Les données acquises avec le capteur SPAD 502 sont un indicateur de la teneur en chlorophylle, elle-même corrélée à la teneur en azote;
c. des mesures d'indice de surface foliaire (ISF ou LAI), décrites à la section 3.3.2.

Les données de travail ont été acquises dans le cadre du projet GEOIDE 2000; une campagne d'échantillonnage a été menée sur un champ test de la ferme MacDonald du campus de l'Université McGill à Ste-Anne-de-Bellevue (détails au chapitre 3).

1.3 Justification de l'approche proposée

Comme énoncé plus haut, l'objectif final du projet est de diagnostiquer le niveau d'azote sur l'ensemble d'un champ de maïs à partir de données hyperspectrales. La relation entre la composition chimique d'une feuille de maïs et son spectre de réflectance est connue [3]. Des travaux effectués dans des laboratoires ont permis d'établir des bases solides. La partie du spectre électromagnétique utile pour établir le diagnostique d'une feuille s'étend du visible à l'infrarouge, tel qu'illustré à la figure 2.

Ces connaissances à l'échelle de la feuille ne peuvent cependant pas être extrapolées simplement à l'échelle de la canopée pour la caractérisation du contenu d'une image numérique hyperspectrale de télédétection. En effet, l'énergie électromagnétique qui parvient au capteur aérien ne se constitue pas seulement de la réponse de la canopée, comme schématisé à la figure 3. Cette énergie reçue par une cellule du capteur provient d'une surface (S) au sol plus ou moins grande (dépendamment du type de capteur et de l'altitude lors de la prise de l'image); de ce fait, elle est la somme des énergies renvoyées par les éléments contenus dans (S) à savoir dans notre cas : les différents types de végétations, le sol apparent, les ombres et les insectes principalement. A cela s'ajoutent les perturbations induites par les caractéristiques de l'air situé entre le capteur et la scène

mais aussi celles de la canopée elle-même (cf. l'indice de surface foliaire présenté par les relations (1-1) et (1-2)).

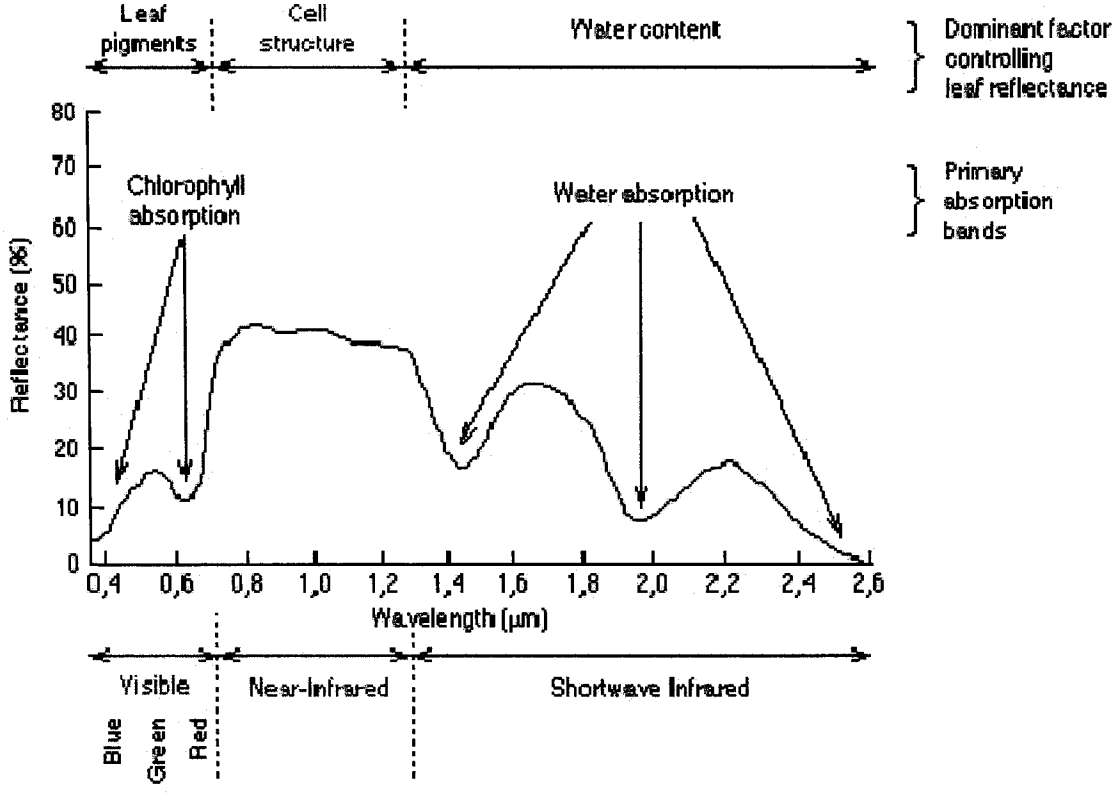


Figure 2 Caractéristiques de réflectance d'une feuille
(P. Lewis & P. Saich, University College London, UK)

Dans le cas idéal d'une image de la canopée d'une seule espèce végétale recouvrant uniformément la totalité de la surface au sol, acquise dans le vide, il serait possible d'étendre les connaissances de l'échelle de la feuille à celle de la zone (S). Dans la réalité, pour les raisons ci-haut énoncées, nous sommes loin de ce cas idéal. Il faut par conséquent trouver un moyen d'isoler l'information pertinente qui concerne la végétation. Une des solutions pour l'étude des couverts végétaux consiste à effectuer des opérations arithmétiques entre des bandes spectrales sélectionnées (cf. ANNEXE 1) pour

amenuiser l'effet de certains paramètres. Les nouvelles « images » ainsi obtenues sont appelées indices de végétation (IV). L'état de l'art sur les IV est présenté dans la revue de littérature au chapitre 2.

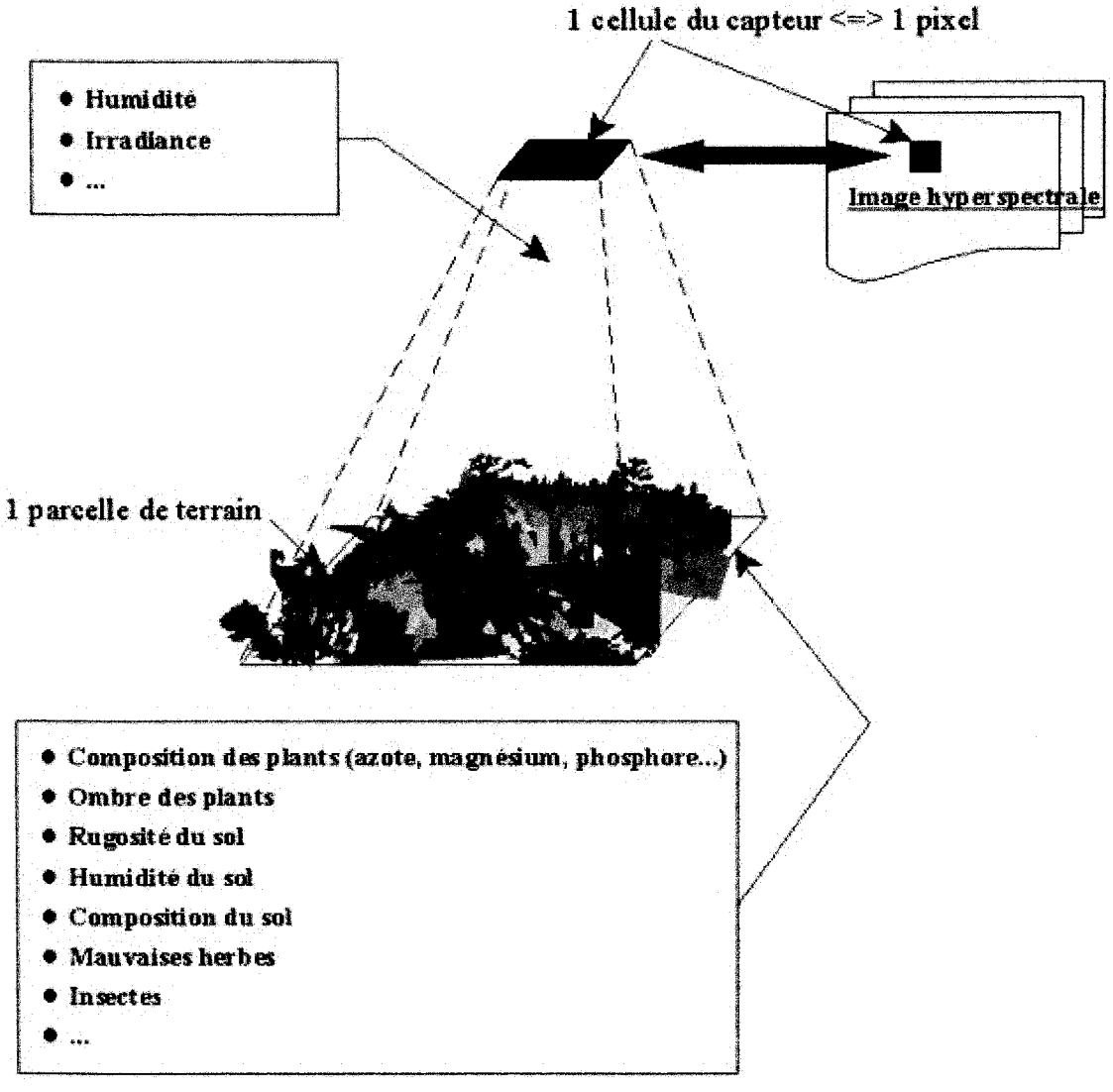


Figure 3 Schéma du contenu d'un pixel d'une image de télédétection

Certains des indices proposés dans la littérature [4-6] permettent d'évaluer l'azote mais ils restent sensibles à d'autres paramètres. Un des plus influents est la densité de

végétation représentée par l'Indice de Surface Foliaire (ISF). Cet indice est défini par les deux relations (1-1) et (1-2) suivantes :

$$ISF = \int_{z=0}^{z=H} u_l(z)dz \qquad (1\text{-}1)$$

avec

$$u_l(z) = N_v(z)A_l \qquad (1\text{-}2)$$

Où : $N_v(z)$ est le nombre de feuilles par m^3 de canopée (continûment intégrable), z est la distance au sol de la couche de la canopée considérée, A_l est la surface moyenne des feuilles et H est la hauteur totale du couvert végétal. De ce fait, on interprète l'ISF comme étant le rapport entre la surface totale des feuilles et la surface au sol; il n'a donc pas de dimension (m^2/m^2). Dans le cas d'un ISF faible, le sol et les tiges des plants sont davantage visibles et la plupart des IV voient leurs performances chuter. Jusqu'à présent, aucun indice proposé n'est totalement indépendant des facteurs qui perturbent l'information pertinente sur la canopée.

En partant du constat précédent, il apparaît difficile d'extraire de l'information précise par la seule étude d'une image de télédétection. Ainsi, l'utilisation de la « réalité du terrain » fournit une information supplémentaire de référence. En télédétection, elle est appelée « ground truth » et elle permet l'utilisation de techniques de classification. En effet, connaissant N mesures d'azote en N points du champ, il est alors possible de séparer ces points en classes de teneur en azote (Faible-Moyen-Élevé par exemple) et d'entraîner un classifieur (KNN par exemple) sur ces données pour classifier par la suite le champ dans sa totalité; ceci est réalisé en utilisant le formalisme présenté en ANNEXE 1. Toutefois, séparer un ensemble de valeurs numériques dans des classes

pose certains problèmes au niveau des données proches des frontières. Pour cette raison, nous ne choisirons pas de technique empruntée à la Reconnaissance de Formes pour valider notre méthode. Au lieu de ça, nous comparons nos résultats aux performances obtenues par des modèles de la littérature, sur notre jeu de données.

1.4 Algorithme proposé

Les IV apportent une réponse pertinente dans de nombreux cas mais aucun n'est robuste pour toutes les configurations réelles possibles. De plus, les IV sont des combinaisons arithmétiques de bandes spectrales. Partant de ces deux constats, l'algorithme proposé a comme objectif de trouver la combinaison de bandes spectrales la plus fortement corrélée à la teneur en azote de la canopée, grâce aux mesures in-situ. Comme il existe un nombre infini de combinaisons de bandes, l'espace des possibles ne peut pas être exploré de façon exhaustive. Pour résoudre ce problème, nous utiliserons la Programmation Génétique qui va permettre, au fil d'itérations successives, de converger vers certaines des combinaisons de bandes représentatives des variations d'azote au sein de la parcelle. L'implémentation de cet algorithme sera réalisée en langage C et sa description sera fournie dans le corps du mémoire.

1.5 Hypothèses de travail

Les hypothèses de notre étude sont :

H1 : il existe une corrélation entre la réflectance de la canopée d'un champ et sa teneur en azote.
H2 : il est possible d'isoler l'emprunte spectrale d'un paramètre biophysique d'un champ en utilisant un algorithme de recherche évolutif.

H3 : la recherche d'indices de végétation par processus évolutif est plus performante que par la méthode classique qui consiste à les établir en raisonnant à partir des propriétés spectrales connues des végétaux.

1.6 Plan de la présentation

La chronologie des étapes pour remplir le mandat est la suivante :

a. apprentissage des notions sur la Géomatique et les SIG;
b. apprentissage du logiciel de SIG PCI GEOMATICA pour le stockage, la manipulation et l'analyse des données géoréférencées de l'étude;
c. création de la base de données géoréférencée;
d. choix des données pertinentes pour l'étude;
e. revue de littérature;
f. élaboration d'une stratégie pour répondre au mandat;
g. apprentissage de la Programmation Génétique comme approche évolutive de l'Intelligence Artificielle;
h. conception de l'algorithme principal en langage C;
i. expérimentations et analyse des résultats;
j. comparaison des résultats obtenus avec ceux des modèles existants;
k. discussion des résultats et suggestions d'améliorations;
l. étude du transfert technologique.

Ces étapes sont détaillées dans les chapitres suivants.

Le chapitre 2 présente une revue de la littérature pertinente à l'étude, soit la place de la télédétection en agriculture de précision et les indices de végétation. Une brève description de la programmation génétique et de ses applications y est aussi présentée.

Le chapitre 3 est consacré aux données de l'étude. Nous présentons toutes les étapes de transformation depuis l'acquisition jusqu'à leur utilisation pour ce travail. Nous en profitons pour décrire brièvement les SIG et le logiciel PCI GEOMATICA.

Au chapitre 4, nous détaillons l'algorithme proposé; nous discutons des choix effectués lors de sa conception et justifions la stratégie de recherche.

Le chapitre 5 est dédié à la présentation et à la discussion des résultats.

Le chapitre 6 est une étude dont le but est de déterminer le nombre minimum de données à acquérir pour permettre un transfert technologique de notre méthode.

Enfin, nous concluons et proposons une série de recommandations visant à l'amélioration de notre contribution.

CHAPITRE 2

REVUE DE LITTÉRATURE

Dans cette revue de littérature, nous présentons tout d'abord la place de la télédétection en agriculture de précision. Nous évaluons son potentiel de développement ainsi que ses limitations. Par la suite, nous présentons les connaissances sur les propriétés spectrales des végétaux; nous faisons un état de l'art sur les indices de végétation (IV) pour la caractérisation de différents stress à différentes échelles. Une attention particulière est portée aux facteurs qui compliquent l'extension des connaissances à la télédétection aérienne. Pour finir, une présentation de la programmation génétique est proposée à travers diverses applications.

2.1 La télédétection en agriculture de précision

La majeure partie de cette section provient d'une revue de littérature proposée par Susan Moran et al. [7] qui synthétise environ 200 ouvrages portant sur l'utilisation, le potentiel de développement et les limitations de la télédétection en agriculture de précision (articles de journaux, papiers de conférences, livres...). Ce travail de synthèse est présenté et complété par des travaux plus récents afin de l'actualiser.

2.1.1 Situation actuelle

L'agriculture de précision a environ une vingtaine d'années. Jusqu'à la fin des années 90, l'utilisation de la télédétection était principalement dédiée à la gestion de l'agriculture à grande échelle pour les applications suivantes [8] :

a. contrôle des surfaces déclarées;
b. occupation du sol;
c. évaluation de l'importance et de l'extension des dégâts lors de catastrophes naturelles (gel, sécheresse, inondations);
d. évaluation de la productivité à grande échelle.

Ce n'est que récemment que les contextes environnemental et économique ont forcé le développement des connaissances pour conduire des stratégies de gestion de la variabilité intraparcellaire par télédétection.

Pour des applications liées à l'agriculture de précision, Moran et al. [7] recensent trois classes d'informations relatives aux caractéristiques des cultures :

a. composantes stables sur la saison (propriétés spatiales du sol, cartes de rendement);
b. composantes variables au cours de la saison (infestation de mauvaises herbes et/ou d'insectes, stress en éléments nutritifs, humidité du sol);
c. informations utiles pour diagnostiquer la ou les raisons de la variabilité de rendement et envisager une stratégie de gestion adéquate.

Dans le cas a, l'imagerie aérienne permet de fournir de l'information pour interpoler des données d'échantillonnage au sol obtenues par des capteurs de rendement. D'après D.S. Long et al., cette technique est la meilleure en comparaison à des méthodes de Krigeage [9]. De même, pour l'interpolation des données sur la composition du sol en début de saison, la télédétection multibande fournit de l'information pertinente [10].

Dans le cas b, des résultats prometteurs ont été obtenus pour établir la variabilité en azote de la canopée pour certaines cultures comme le riz, le blé et le maïs [11], au moyen de données hyperspectrales. Pour établir les niveaux d'infestation de la folle-

avoine au sein des cultures de blé, l'utilisation de l'imagerie aérienne a été démontrée pertinente en terme de réactivité, de coût et de précision [12].

Dans le cas c, l'enjeu est de trouver les causes de la variabilité spatiale du rendement afin de développer une stratégie de gestion adaptée. Des systèmes experts existent qui permettent de déterminer des relations cause/effet; certains de ces modèles intègrent des données de télédétection pour l'analyse et la détermination de ces relations [13]. Ici aussi, les auteurs Moran et al. [7] prévoient un fort développement pour les données de télédétection.

Aujourd'hui, de grands groupes comme SPOT[3] proposent des solutions (FARMSTAR[4]) pour la gestion d'intrants comme l'azote sur certaines cultures comme le blé, à partir d'images provenant de leurs satellites. En France, en 2004, environ 200 000 hectares de cultures ont été assistés par des solutions de FARMSTAR; Sur cette période, les bénéfices moyens enregistrés sont les suivants :

a. une augmentation de la marge brute (de 45$/ha à 140$/ha selon le type de culture);
b. un gain de temps par la réduction de l'échantillonnage;
c. la suppression d'un 3ème apport d'azote dans presque 30% des cas;
d. une amélioration de la qualité globale des produits;
e. une diminution de l'impact sur l'environnement, approuvée par des organismes nationaux (ministère de l'agriculture...).

Le potentiel est donc énorme puisqu'il reste encore de nombreuses lacunes dans l'analyse des données de télédétection. Dans la section suivante, nous détaillons les champs d'application pour cette technologie relativement aux besoins de l'agriculture.

[3] http://www.spotimage.fr
[4] http://www.farmstar-space.com/

2.1.2 Opportunités de développement

Le potentiel de développement de la télédétection en agriculture de précision concerne huit points spécifiques [7]. Cette classification jugée pertinente est présentée ci-dessous :

a. définition des unités pour la conduite précise des cultures en combinant des données de terrain (sol et canopée) avec des données hyperspectrales de télédétection au moyen de techniques d'interpolation;
b. cartographie du rendement au moyen d'images multispectrales obtenues à un stade avancé de la croissance des plants. Le couplage avec des données issues de modèles agrométéorologiques est également envisageable;
c. cartographie de la variabilité du sol en nutriments, en début de saison lorsqu'il n'y a pas encore de végétation, grâce à l'imagerie multibande;
d. détermination des huit catégories de caractéristiques variables au cours de la saison :

   - humidité du sol;
   - stade phénologique de la culture;
   - biomasse et rendement;
   - taux d'évapotranspiration;
   - déficiences en nutriments (azote, phosphore…);
   - maladies végétales;
   - infestations de mauvaises herbes;
   - infestations d'insectes.

e. fourniture des données d'entrée pour des systèmes d'aide à la décision (DSS, « Decision Support System ») ou de base de connaissances pour des systèmes experts dans le but de comprendre les origines de la variabilité spatiale de production. Il est également envisageable de coupler ces données à des modèles

agrométéorologiques afin de trouver les causes de la variabilité de production et celles de l'hétérogénéité de composition du sol;

f. données multispectrales de résolution spatiale grossière utiles comme informations sur les paramètres météorologiques et climatiques comme l'insolation globale, les pluies ou encore le rayonnement photosynthétiquement actif (RAP ou PAR). En effet, les données des stations météorologiques ne reflètent pas précisément la variabilité des conditions sur la zone qu'elles sont censées couvrir;
g. production de modèles numériques de terrain (MNT ou DEM pour « Digital Elevation Model ») à partir d'un couple d'images stéréoscopiques, permettant la connaissance précise de la topographie du terrain;
h. détermination des dates critiques d'intervention en cas de dommages (maladies, infestations, catastrophes naturelles...) et mesure de l'étendue des dégâts grâce à l'imagerie multibande.

Pour de plus amples informations sur ces points, le lecteur pourra se référer aux ouvrages placés en référence de la revue détaillée dans cette section.

2.1.3 Limitations

Les premières limites de l'agriculture de précision sont le prix encore excessif des images de télédétection et également l'accès difficile à cette ressource; ces freins au développement sont en perpétuel retrait grâce notamment à la démocratisation des procédés d'acquisition.

L'autre catégorie de limitations concerne l'acquisition, les prétraitements (corrections) et les traitements des données de télédétection. Le succès de la télédétection repose, pour la majorité des applications précitées, sur la précision de l'information quantitative qu'elle renferme. De ce fait, l'étape de calibrage d'un capteur peut être critique, principalement pour la télédétection aéroportée. Une fois acquises, les données doivent subir une

correction radiométrique afin d'éliminer les effets de l'atmosphère. Des améliorations dans cette étape critique sont encore à apporter pour gagner en précision. D'autres effets comme le vignetage (artefact optique qui diminue la luminosité sur les bords de l'image) viennent perturber le signal reçu et doivent être corrigés lors de la phase de prétraitement. Les prétraitements concernant les rectifications géométriques sont maîtrisés et ne sont pas une source importante d'erreur s'ils sont effectués correctement. Le couvert nuageux est également un frein pour la fiabilité de l'imagerie satellite puisqu'il peut rendre inexploitable une image à un moment critique du processus agricole.

Les améliorations dans ces phases de prétraitement sont primordiales pour fournir une base solide à l'analyse pour l'extraction d'information de ces données; les investigations (comme notre étude) venant en aval de ces prétraitements, leur succès en dépend fortement.

2.2 Propriétés des feuilles

2.2.1 Structure des feuilles et modèles de réflectance

Les propriétés optiques des feuilles dépendent en grande partie de leur structure montrée à la figure 4. Selon les caractéristiques de la cuticule, la réflexion spéculaire de la feuille va varier. L'épaisseur de la feuille ainsi que sa composition interne vont également influencer l'absorption et la transmission de l'énergie aux différentes longueurs d'onde.

Des modèles de simulation de réflectance de canopée ont été développés [14-17] pour permettre de générer des données virtuelles de réflectance, pour des paramètres variables de l'ISF, de position du soleil, de distribution angulaire des feuilles LAD (« Leaf Angle Distribution ») ou encore de type de sol sous-jacent. Plusieurs approches ont été utilisées

pour modéliser les propriétés optiques des feuilles[5] parmi lesquelles on peut nommer le lancer de rayons, les modèles à N flux et les modèles à empilement de couches [16]. Ces modèles considèrent les propriétés spectrales de chacune des couches constitutives des végétaux (cf. figure 4) et les assemblent pour déterminer les proportions qui sont transmises, absorbées et réfléchies. Un modèle stochastique faisant intervenir les chaînes de Markov a récemment été repris par Maier et al [18]. Dans ce modèle, il y a quatre états possibles pour le rayonnement (incident, réfléchi, transmis ou absorbé); des probabilités sont définies pour la transition d'un état à un autre, lors du passage du rayon au travers des différentes couches de la feuille (cf. figure 4).

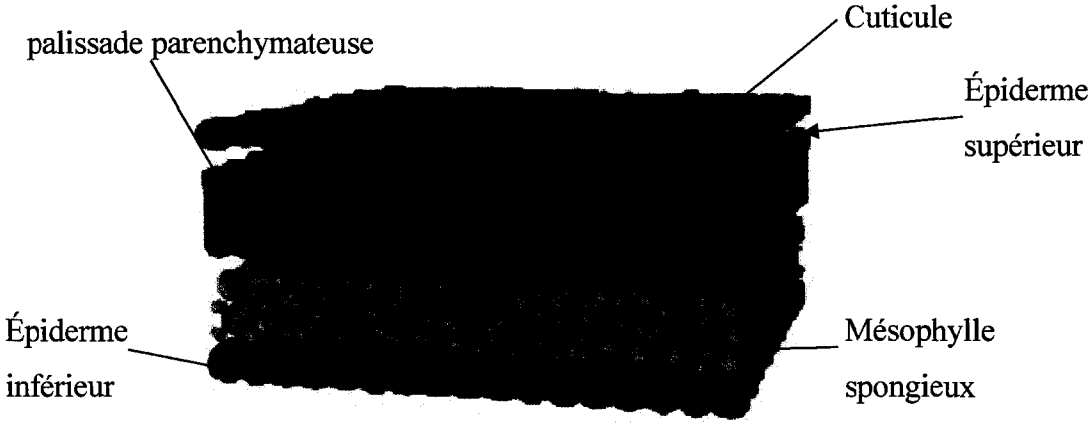


Figure 4 Structure d'une feuille
(University of London)

Au milieu des années 1990, des couplages de modèles optiques de feuilles avec des modèles de canopées sont apparus [19, 20]. Jacquemoud et al. proposent un comparatif de quatre de ces modèles de canopée couplé avec le modèle foliaire PROSPECT [14]. Pour des informations supplémentaires, le lecteur pourra se référer aux documents cités, ce domaine n'étant pas l'objet de notre étude.

[5] http://www.sigu7.jussieu.fr/Led/LED_leafmod.htm

2.2.2 Caractéristiques spectrales des feuilles : études en laboratoire

Les études menées sur le spectre de réflectance des feuilles ont rapporté des correspondances de certaines longueurs d'ondes λ avec des caractéristiques physiologiques. Josep Peñuelas et al. [21] rappellent ces relations dans un article de synthèse :

a. $\lambda = 430$ nm et $\lambda = 445$ nm pour la teneur en caroténoïdes;
b. $\lambda = 531$ nm et $\lambda = 570$ nm pour la teneur en xanthophylles;
c. $\lambda = 550$ nm à $\lambda = 675$ nm et position de la « red edge » (définie à la section 2.4.1) pour la concentration en chlorophylle;
d. $\lambda = 700$ nm à $\lambda = 800$ nm pour la teneur en pigments bruns;
e. $\lambda = 800$ nm et $\lambda = 900$ nm pour la caractérisation de la structure de la feuille;
f. $\lambda = 970$ nm pour la teneur en eau;
g. $\lambda = 800$ nm à $\lambda = 900$ nm et $\lambda = 680$ nm pour la biomasse verte (masse totale des éléments chlorophylliens).

Ces résultats proviennent d'études menées en laboratoire dans des environnements contrôlés. Comme nous l'avons mentionné dans l'introduction à la section 1.3, lors du passage vers la télédétection à l'échelle de la canopée en environnement naturel, des facteurs viennent influencer la réflectance des couverts végétaux; de ce fait, ces résultats ne sont plus utilisables directement [22]. Ces facteurs sont détaillés dans le paragraphe suivant.

2.3 Effets influençant la réflectance des couverts végétaux

Trois types de facteurs viennent influencer la réflectance d'un couvert végétal; les facteurs externes, les facteurs propres à la végétation et ceux relatifs au sol [22].

2.3.1 Facteurs externes

Le premier facteur externe est la nature du rayonnement incident. Il se compose d'un rayonnement direct et d'un rayonnement diffus. Cette fraction du rayonnement diffus dépend principalement de la nébulosité et de façon générale, la réflectance du couvert végétal est peu sensible à ce facteur [17].

Le deuxième facteur externe est relié à la géométrie de la mesure. Il faut s'assurer en premier lieu que la surface échantillonnée est représentative du couvert végétal observé pour minimiser la variabilité locale non significative (diamètre minimal de 1 mètre pour le maïs). D'autre part, les couverts végétaux ne sont pas des surfaces lambertiennes (ce qui revient à dire qu'ils ne renvoient pas l'énergie lumineuse également dans toutes les directions) et les positions relatives du soleil et du capteur ont donc une influence sur leurs valeurs de réflectance. Le phénomène le plus connu est celui du « point chaud » (connu en anglais par le terme « hot spot ») qui peut être décrit à partir de la figure 5; il apparaît lorsque le capteur se situe dans le plan principal du soleil ($\psi_0 = 0$) avec un angle zénithal solaire égal (ou proche) de l'angle zénithal de visée de capteur ($\theta_s \approx \theta_0$).

Le dernier facteur externe est l'influence de l'atmosphère. Les deux mécanismes qui viennent perturber le rayonnement sont l'absorption et la diffusion. L'absorption entraîne la disparition de photons absorbés par les gaz et la vapeur d'eau; elle dépend de la longueur d'onde. La diffusion entraîne la redistribution spatiale des photons qui heurtent les molécules (gaz ou vapeurs) ce qui a pour effet de diminuer le signal dans la direction du rayonnement incident.

Des indices de végétation ont été élaborés pour minimiser les effets perturbateurs de l'atmosphère et certains seront présentés à la section 2.4.5 [23, 24].

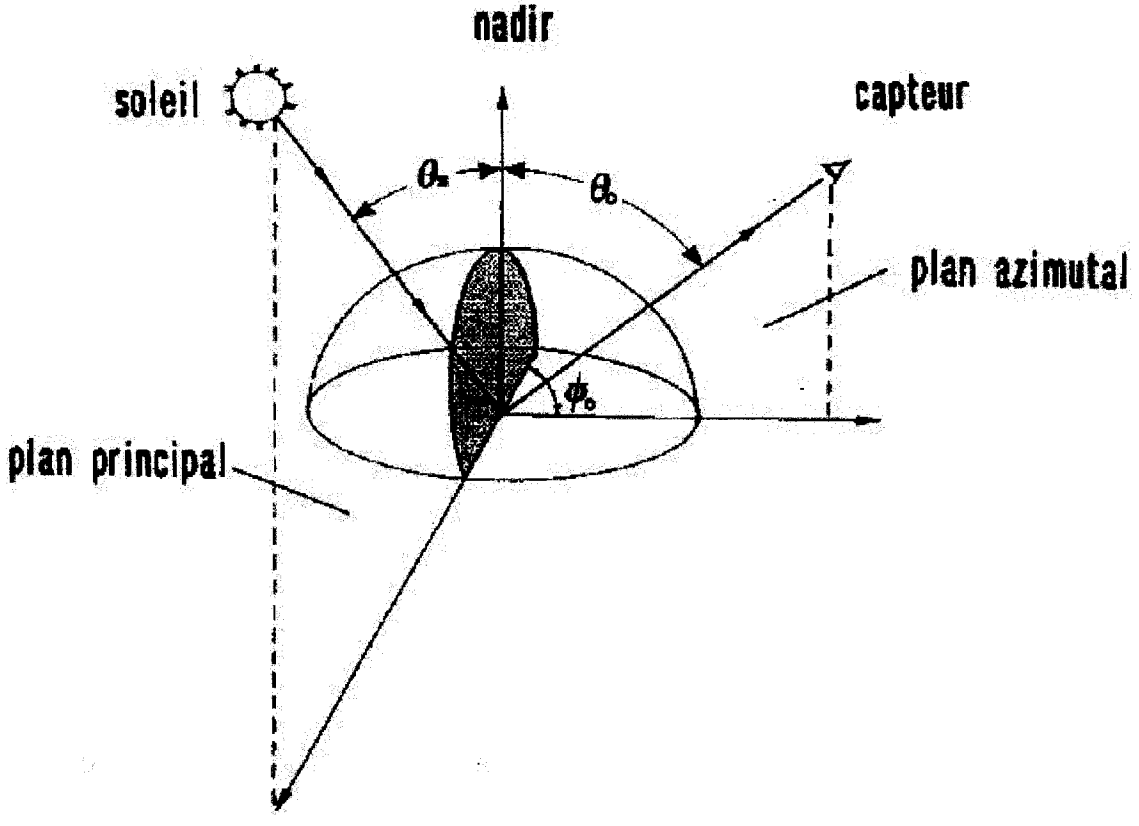


Figure 5 Géométrie de mesure d'un capteur aérien [22]

2.3.2 Facteurs liés à la végétation

Les feuilles constituent la majeure partie de la surface d'un couvert végétal. Les autres éléments comme les tiges, les fruits ou encore les fleurs sont rarement pris en compte même si leur présence influence la réflectance pendant une partie de l'année. La composition biochimique gouverne une partie du spectre de réflectance, comme il a déjà été mentionné auparavant (lignine, cellulose, protéines, pigments…).

La réflectance du couvert végétal est également influencée par son architecture. Les paramètres qui permettent de la définir sont :

a. la distribution spatiale de la végétation au sol;
b. l'indice de surface foliaire (ISF ou LAI);
c. l'inclinaison des feuilles caractérisée par la fonction de distribution angulaire LAD.

La distribution spatiale dépend principalement du type de culture observé, de la configuration de plantation choisie par l'agriculteur et du stade de maturité. Lorsque la végétation s'étoffe au fil de la saison, ce facteur devient moins important.

L'indice ISF est un des facteurs les plus influents. En agriculture, il varie généralement de 0 (pour un sol nu) à 8-10 pour une culture très fournie [22]. La réflectance varie également en fonction de cet indice et son comportement dépend de la longueur d'onde comme montré à la figure 6.

Enfin, les feuilles ne sont pas inclinées uniformément au sein d'un couvert végétal. Il faut par conséquent définir une inclinaison moyenne ainsi qu'une fonction de distribution LAD. Généralement, c'est une fonction de distribution sphérique qui est considérée dans les modèles [25, 26].

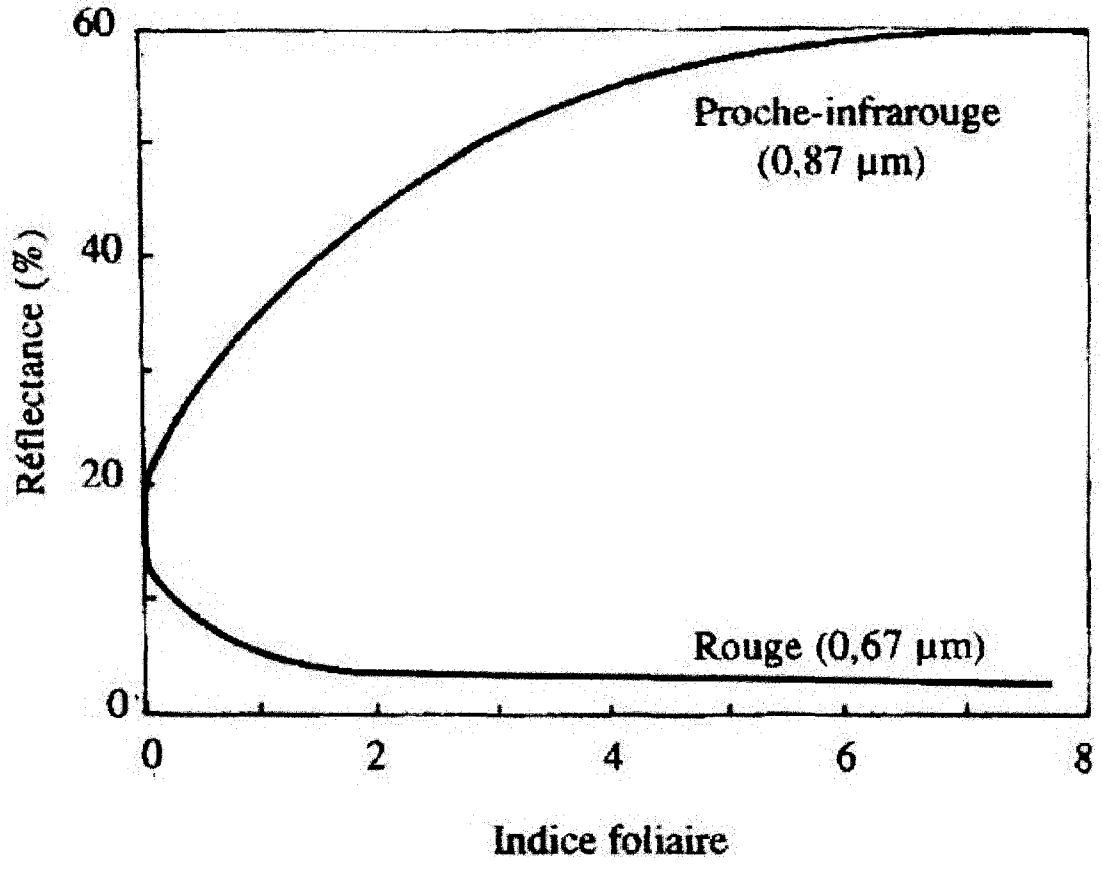


Figure 6 Réflectance d'un couvert végétal en fonction de l'ISF [22]

2.3.3 Facteurs liés au sol

Lorsque le sol est apparent sous la végétation, tel que pour des ISF faibles, ses caractéristiques spectrales vont intervenir. Sa texture (rugosité) et composition (teneur en matière organique, en eau, en air...) sont les deux principales caractéristiques qui vont modifier le spectre de réflectance.

Nous discuterons par la suite des indices qui ont été élaborés pour minimiser les effets du sol sous-jacent des couverts végétaux.

2.4 Indices de Végétation (IV)

2.4.1 Principaux indices

Les indices pour l'évaluation des conditions physiologiques des plantes ont été élaborés au moyen de mesures hyperspectrales effectuées en laboratoire et au sol. Les caractéristiques des principaux indices sont rappelées dans l'article de Peñuelas et Filella [21] et sont décrites ci-dessous :

a. Simple Ratio : SR := PIR/R;
b. NDVI := (PIR-R) / (PIR+R) := $(R_{800} - R_{670}) / (R_{800}+R_{670})$;
c. λ_{RE} := longueur d'onde de la pente maximum de l'augmentation de la réflectance lors du passage du rouge à l'IR (« red edge »);
d. PRI := $(R_{531}-R_{570})/(R_{531}+R_{570})$;
e. SIPI := $(R_{800}-R_{445})/(R_{531}-R_{680})$;
f. WI := R_{900}/R_{970};
g. NPQI := $(R_{415}-R_{435})/(R_{415}+R_{435})$.

Où PIR est la réflectance dans le Proche Infrarouge et R_i représente la réflectance à la longueur d'onde *i* exprimée en nanomètres (nm). Les indices et leur abréviation sont définis ci-après.

- Le SR est principalement utilisé pour estimer la biomasse verte [27].
- Le NDVI (Normalized Difference Vegetation Index), appelé aussi indice d'activité végétale ou encore indice de Tucker, est utilisé pour l'évaluation de la biomasse verte. Il est également employé pour estimer l'ISF de couverts végétaux. D'autres utilisations ainsi que sa forme généralisée seront discutées à la sous-division intitulée « **indices de différence normalisée** ».

- Le λ_{RE} (longueur d'onde de la « Red Edge »), mis en évidence à la figure 7, est défini comme la longueur d'onde pour laquelle la dérivée première du spectre de réflectance s'annule (point d'inflexion sur le spectre de réflectance), lors de la transition du rouge à l'infrarouge. C'est un indicateur de la teneur en chlorophylle de la feuille.
- Le PRI (Photochemical Reflectance Index) est un indicateur de la fluorescence et sera présenté dans la section consacrée à la description de ce phénomène.
- Lorsque les feuilles sont sujettes à des carences en éléments nutritifs, la teneur en caroténoïdes augmente au détriment de celle en chlorophylle. Ainsi, un indice comme le SIPI (Structure Independent Pigment Index) est un bon indicateur du ratio des pigments caroténoïdes/chlorophylle pour certaines plantes comme le tournesol [28].

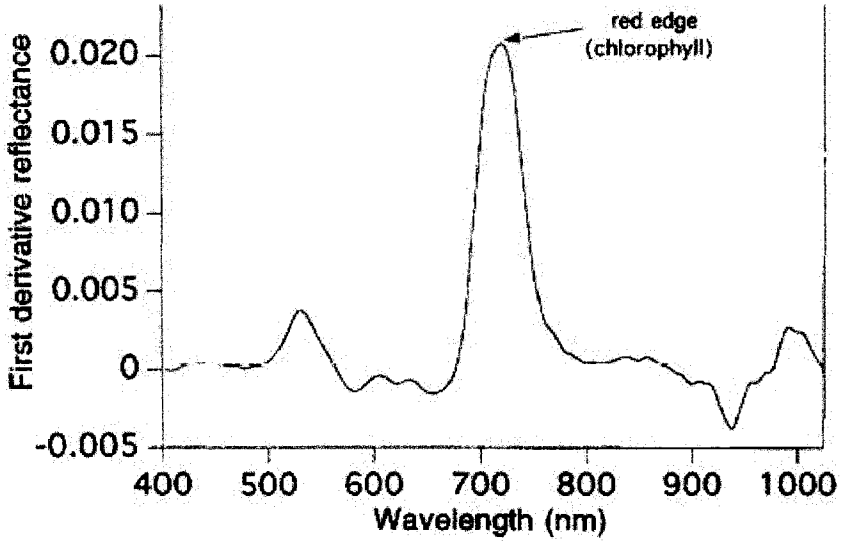


Figure 7 1ère dérivée du spectre de réflectance / Position de la « red edge » [21]

- Le WI (Water Index) [29], est décrit comme un indicateur de la concentration en eau de la plante (Plant Water Content, PWC). En calculant le ratio WI/NDVI, on corrige

l'effet du NDVI sur le WI puisqu'il existe généralement une corrélation entre la biomasse verte et la teneur en eau.

- Le NPQI donne un indicateur de la dégradation de la chlorophylle [30] mais cet indice de différence normalisée est moins connu et moins utilisé que ceux précités.

Le problème de ces indices est que lorsqu'on passe à l'échelle de la canopée (télédétection aérienne), ils sont perturbés par les influences de la réflectance du sol sous-jacent et de l'ISF. La position du soleil, l'hétérogénéité de la canopée, le type de végétation, sa structure ainsi que les interférences de l'atmosphère affectent également leurs performances.

2.4.2 Indices de différence normalisée

Concernant l'évaluation de la biomasse qui est un paramètre très intéressant du point de vue de l'agriculture, pour prévoir les récoltes par exemple, il a été montré par Gamon et al. [27] que le NDVI était un bon indicateur de ce paramètre. Cet indice est le plus utilisé puisqu'il permet d'évaluer beaucoup d'autres paramètres comme par exemple le niveau d'infestation d'acariose pour certaines cultures [31]. Il est également relié dans certains cas à la structure de la canopée ainsi qu'à son activité photosynthétique [27]; de cette étude menée sur trois types de végétation en Californie, il est ressorti que le NDVI est un bon indicateur de la biomasse verte, de la teneur en chlorophylle et de la teneur en azote des feuilles dans le cas de végétation éparse ($0 < ISF < 2$). Dans le cas de couverts plus denses ($ISF > 2$), ce n'est plus le cas et son couplage avec l'indice SR ne donne pas de résultats significatifs.

Hansen et al. [32] ont eu l'idée de tester plusieurs combinaisons possibles de bandes du type NDVI, définis par l'équation générale (2-1).

$$NDVI(\lambda 1, \lambda 2) = \frac{\lambda 1 - \lambda 2}{\lambda 1 + \lambda 2} \quad (2\text{-}1)$$

Cette étude a été faite sur une culture de blé à partir des données d'un spectromètre manuel terrestre (télédétection au sol), sur l'intervalle [400 ; 900] nm. Les auteurs ont cherché le NDVI ($\lambda 1$,$\lambda 2$) le plus fortement corrélé aux mesures de chlorophylle (concentration Chl_{conc} et quantité $Chl_{densité}$), aux mesures d'azote (concentration N_{conc} et quantité $N_{densité}$), aux mesures de biomasse verte (GBM) ainsi qu'à l'ISF. Le plus souvent, les meilleurs modèles de régression sont de type exponentiel. Pour chacune de ces expériences, les couples ($\lambda 1$,$\lambda 2$) les plus pertinents sont :

a. GBM : $\lambda 1 = 565$ nm et $\lambda 2 = 708$ nm;
b. ISF :
    - $\lambda 1 = 711$ nm et $\lambda 2 = 720$ nm;
    - $\lambda 1 = 746$ nm et $\lambda 2 = 750$ nm;
    - $\lambda 1 = 556$ nm et $\lambda 2 = 730$ nm;
    - $\lambda 1 = 556$ nm et $\lambda 2 = 760$ nm;
c. Chl_{conc} : pas de résultat intéressant;
d. $Chl_{densité}$: $\lambda 1 = 717$ nm et $\lambda 2 = 732$ nm;
e. N_{conc} : $\lambda 1 = 565$ nm et $\lambda 2 = 708$ nm;
f. $N_{densité}$:
    - $\lambda 1 = 730$ nm et $\lambda 2 = 759$ nm;
    - $\lambda 1 = 717$ nm et $\lambda 2 = 770$ nm;
    - $\lambda 1 = 720$ nm et $\lambda 2 = 839$ nm.

Finalement, Hansen et al. obtiennent comme meilleur modèle de corrélation pour la densité d'azote un coefficient de détermination $R^2 = 69\%$. Pour ce qui est de la concentration en azote, ils arrivent à $R^2 = 56\%$. Le meilleur résultat est pour la prévision

de la biomasse verte puisque $R^2 = 84\%$. Cette étude a confirmé également ce qui est déjà bien connu à savoir que le NDVI est corrélé fortement à l'ISF; c'est d'ailleurs un résultat qui peut être utilisé pour évaluer l'ISF lorsque cette donnée n'est pas mesurée directement en champ. D'une façon générale, les résultats de cette étude montrent que cette famille d'indices ne peut pas répondre à l'évaluation de tous les paramètres caractéristiques d'une grande culture. La nécessité d'explorer d'autres formes d'indices et d'autres approches est par conséquent primordiale.

L'indice NDVI reste malgré tout utilisé pour l'évaluation de l'ISF mais il sature lorsque la végétation devient trop dense (cf. figure 2.a [33]). De ce fait, l'indice « renormalisé » RDVI [34] et le simple ratio modifié MSR [35] ont été proposés et sont définis par les relations (2-2) et (2-3) suivantes.

$$RDVI = \frac{\left(R_{800} - R_{670}\right)}{\sqrt{R_{800} + R_{670}}} \tag{2-2}$$

$$MSR = \frac{\frac{R_{800}}{R_{670}} - 1}{\sqrt{\frac{R_{800}}{R_{670}} + 1}} \tag{2-3}$$

RDVI se comporte de façon plus linéaire en fonction de l'ISF et MSR est plus sensible aux variations des paramètres biochimiques de la végétation. D'autres indices sont présentés par Haboudane et al. [25], toujours dans l'esprit de minimiser l'effet de certains paramètres pour permettre l'évaluation d'autres. Certains de ces indices seront discutés dans la section consacrée aux indices ajustés.

Nous le voyons, les connaissances sont assez bien établies lors de l'observation du spectre de réflectance d'une feuille ou bien de celle de mesures au sol sur la canopée. Le passage à l'échelle de la télédétection aérienne reste problématique en raison des multiples facteurs qui viennent perturber le signal. Pour cette raison, des indices ont été développés pour minimiser les effets de paramètres comme le sol ou l'atmosphère.

2.4.3 Évaluation de la chlorophylle

La concentration en chlorophylle est fortement corrélée à celle en azote. Pour de faibles concentrations, c'est la région autour de 675 nm (maximum d'absorption) qui est la plus sensible aux variations; pour des concentrations moyenne à forte, c'est à la longueur d'onde 550 nm que la sensibilité est meilleure [16].

Pour certaines cultures comme celle du blé, des indices de « concentration de chlorophylle » ont été développés [36] qui permettent d'estimer par la suite la teneur en azote.

L'apparition d'investigations sur la première dérivée du spectre de réflectance a permis de mettre en relief la « red edge » (cf. figure 7); c'est la longueur d'onde à laquelle la pente entre la réflectance du rouge et de l'infrarouge est maximum. À l'échelle de la feuille, sa position et sa forme sont un bon indicateur de la teneur en chlorophylle mais à l'échelle de la canopée, ce résultat est plus controversé. Toutefois, les imageurs hyperspectraux permettant d'obtenir de meilleures précisions, des recherches sont menées pour tenter d'évaluer le plus précisément possible la position de cette longueur d'onde [37].

Des indices de végétation ont été développés pour évaluer l'absorption par les pigments chlorophylliens. Le CARI [38] puis le MCARI [26] (2-4) ont été conçus dans ce but.

$$MCARI = \left[(R_{700} - R_{670}) - 0.2(R_{700} - R_{550})\right](R_{700} / R_{670}) \quad (2\text{-}4)$$

Il est cependant apparu que MCARI était sensible aux variations de l'ISF à 60% et seulement à 27% aux variations de chlorophylle. C'est donc principalement pour l'ISF que cet indice peut être utilisé en fin de compte.

Avec le même objectif initial d'évaluer l'énergie radiative absorbée par les pigments foliaires, Broge et Leblanc [6] ont proposé le TVI (2-5).

$$TVI = 0.5\left[120(R_{750} - R_{550}) - 200(R_{670} - R_{550})\right] \quad (2\text{-}5)$$

Cet indice utilise donc la longueur d'onde dans le vert au pic de réflectance, au minimum de réflectance dans le rouge et dans le début du proche infrarouge (PIR ou NIR). L'idée est que l'aire du triangle formé par ces trois points du spectre croît lorsque l'absorption chlorophyllienne augmente; en effet, cette absorption des pigments chlorophylliens entraîne une diminution de l'énergie réfléchie dans le rouge et une augmentation des tissus foliaires [6] entraînant une hausse de la réflectance dans le PIR. Cependant, Haboudane et al. [25] nuancent ces résultats en précisant qu'il existe des effets indirects qui viennent contrebalancer ce « principe du triangle ».

2.4.4 Fluorescence

La fluorescence est largement utilisée pour établir les conditions physiologiques des plantes [39]. En effet, une partie de l'énergie absorbée par la feuille, pour effectuer la photosynthèse, est dissipée sous forme de chaleur tandis qu'une autre l'est sous forme de fluorescence; ces paramètres sont donc un indicateur de l'activité de la plante. Des

indices de réflectance ont été développés pour évaluer la santé des plantes grâce à la mesure de ces deux types d'émissions. À l'échelle de la feuille ainsi qu'à celle de la télédétection au sol, ces indices comme le PRI (corrélé à la variation de fluorescence) [21] sont largement utilisés et donnent de bonnes estimations de l'activité chlorophyllienne et donc indirectement de la santé des plantes. Cependant, à l'échelle de la canopée du champ, ces émissions étant faibles, il est dangereux de tirer des conclusions du fait des perturbations apportées par un grand nombre de paramètres comme la structure de la canopée, l'hétérogénéité du paysage, les erreurs de calibrage ou encore les interférences atmosphériques.

Une étude menée par Pablo J. Zarco-Tejada et al. [40, 41] a permis de conclure que l'estimation des paramètres biochimiques de la canopée étaient possibles par télédétection sur des forêts d'érables. De bonnes corrélations entre les données de terrain et celles de laboratoire ont été trouvées; les auteurs laissent entendre que l'approfondissement des investigations menées sur la fluorescence permettrait d'améliorer l'estimation de certains paramètres biophysiques comme la teneur en pigments chlorophylliens.

2.4.5 Indices ajustés pour la télédétection aérienne

Pour tenter d'éliminer les effets du sol et de l'atmosphère sur la réflectance des couverts végétaux, des indices ajustés ont été proposés par certains auteurs. Huete [42] est le premier à proposer le SAVI (2-6) qui vise à réduire l'influence du sol sous les plants.

$$SAVI = \frac{(1+L)(R_{800} - R_{670})}{(R_{800} + R_{670} + L)} \qquad (2\text{-}6)$$

La valeur de L est une fonction de la densité de végétation et sa détermination requiert la connaissance à priori de la quantité de végétation. La valeur proposée par Huete est L = 0.5 qui convient en première approximation à tous les types de sol. Cependant, Qi et al. [43] ont tenté d'améliorer SAVI en proposant MSAVI défini par la relation (2-7).

$$MSAVI = \frac{1}{2} * [2 * R_{800} + 1 - \sqrt{(2 * R_{800} + 1) - 8(R_{800} - R_{670})}] \qquad (2\text{-}7)$$

Ce nouvel indice ne fait plus apparaître de facteur L. L'étude récente de Broge et Leblanc [6] a permis de faire ressortir MSAVI comme le meilleur estimateur de l'ISF, et ce même pour des canopées très denses (ISF élevés).

Pour minimiser les effets de l'atmosphère, Kaufman et Tanre [23] ont remplacé la réflectance du rouge R_r utilisée dans le NDVI et dans le SAVI par une combinaison du rouge et du bleu R_{rb} (2-8).

$$R_{rb} = R_r - g(R_b - R_r) \qquad (2\text{-}8)$$

Cela permet alors de corriger l'absorption par les aérosols dans la zone du rouge. Ainsi, avec la valeur g = 1 recommandée par les auteurs, ils proposent le SARVI donné par l'équation (2-9) suivante :

$$SARVI = \frac{(1 + L)(R_{800} - R_{rb})}{(R_{800} + R_{rb} + L)} \qquad (2\text{-}9)$$

Cet indice minimise donc à la fois les effets de sol et les effets atmosphériques. D'autres améliorations ont été proposées par la suite, principalement pour réduire les effets des nuages sur les images-satellites (SARVI2) [24].

2.5 Discussion sur les propriétés spectrales

Nous l'avons vu, de nombreuses investigations ont été menées autour de l'analyse des propriétés spectrales des végétaux. De nombreux IV ont été proposés par des chercheurs dont les principaux ont été présentés. Malgré leur aspect empirique et intuitif, certains comme les indices de différence normalisée ont une justification mathématique [44]. Un point saillant de cette revue de littérature est la difficulté de généralisation des propriétés spectrales des couverts végétaux à l'échelle de la canopée; en effet, de nombreuses causes peuvent avoir les mêmes conséquences sur les spectres de réflectance acquis par voie aérienne; la combinaison de plusieurs stress à divers niveaux d'amplitude (sécheresse, carence en azote,...), associé aux bruits créés par l'atmosphère et le sol, peuvent entraîner des réflectances identiques à des longueurs d'onde données et par conséquent empêcher un diagnostique précis à partir d'une longueur d'onde. Combiner des bandes spectrales entre elles permet d'apporter des réponses dans quelques cas, mais il semble difficile d'établir des indices permettant de répondre à toutes les configurations possibles. Dans sa thèse, Bannari [45] parvient aux mêmes conclusions à savoir que, pour son application spécifique, aucun indice de végétation ne permet d'être indépendant de l'ensemble des effets extérieurs; en d'autres termes, lorsqu'un indice est résistant à un effet extérieur, il devient sensible à d'autres.

Pour les applications en agriculture de précision, il y a donc la nécessité de découvrir de nouveaux indices pour palier à ce genre de problème [21] ou d'en utiliser plusieurs pour une prise de décision. Toutefois, en raison de la variabilité dans les spectres de réflectance induite par les facteurs mentionnés plus tôt, il apparaît difficile pour le moment de s'extraire des données de terrain (« ground truth »). L'acquisition de données

au sol couplée à celle de données aériennes reste probablement la solution pour permettre une classification précise de la variabilité au sein des cultures. C'est cette approche que nous nous proposons d'utiliser dans notre étude, et ce à l'aide de la programmation génétique.

2.6 La programmation génétique (PG)

Dans un premier temps, les origines de la PG ainsi que quelques unes de ses applications seront présentées. Par la suite, nous ferons une description générale de cette méthode pour permettre au lecteur d'apprécier la démarche proposée dans le chapitre consacré à notre algorithme; cette partie se veut simplement informative sur le sujet et en aucun cas exhaustive.

2.6.1 Origines et applications

La programmation génétique (PG) fait partie de l'apprentissage machine (Machine Learning, ML) et plus précisément de la branche des algorithmes évolutionnaires en intelligence artificielle. La PG, proposée par J.R. Koza en 1992 [46] est issue des algorithmes génétiques (AG ou GA) dont John Holland est le père-fondateur en 1975 [47, 48]. La principale différence avec les AG réside dans le fait que la forme et la taille de la solution recherchée en PG n'est pas connue; seuls la grammaire et l'alphabet de cette solution le sont.

Les fondements de la PG sont inspirés de la théorie de l'évolution de Darwin [49]; on y retrouve le principe général de sélection des meilleurs « individus » (solutions potentielles) d'une « population » (sous-ensemble de solutions possibles), en terme d'adaptation aux contraintes de leur « milieu » (problème étudié). Les détails sur le principe de la PG sont donnés dans la section suivante.

L'objectif de la PG est de chercher la solution globale pour une application donnée. Elle est recommandée pour les problèmes suivants :

a. problèmes non-résolus de façon analytique;
b. problèmes multicritères (i.e. plusieurs objectifs ou contraintes à satisfaire);
c. problèmes mal posés;
d. problèmes complexes;
e. problèmes dont l'espace des possibles est de dimension infinie;
f. couplage avec des méthodes locales.

D'une façon générale, cette technique convient pour la résolution de problèmes dont on ne connaît pas le processus mais dont on est capable d'évaluer la performance de solutions candidates.

On retrouve la PG dans de nombreux domaines pour l'optimisation de problèmes ayant les caractéristiques ci-haut mentionnées. En informatique, cet outil est très utilisé pour faire évoluer des programmes [50]. On retrouve aussi la PG pour la « résolution » de problèmes mathématiques complexes [51], en électronique pour optimiser la réalisation de circuits [52], en biologie moléculaire [53] pour diverses applications de classification, en économie [54] et dans bien d'autres domaines encore. Pour ce qui est de l'étude d'images multibandes par PG, on retrouve assez peu de travaux [55, 56]. Dans le domaine de l'agriculture de précision, quelques applications ont été proposées utilisant des réseaux de neurones appliqués à l'imagerie [57] ; MengBo Li et al. proposent l'utilisation de techniques d'intelligence artificielle pour la résolution d'un problème multicritère (minimisation du nitrate dans les eaux de ruissellement, maximisation de la production et maximisation des profits) relié à l'application de fertilisants azotés sur un champ de maïs [58]. Il apparaît que les bons résultats des techniques évolutives et autres outils empruntés à l'intelligence artificielle suscitent une certaine curiosité; ceci a pour

effet d'accroître le nombre de leur utilisation dans des applications reliées à l'agriculture de précision.

2.6.2 Description générale de la PG

Le schéma de principe de la PG est présenté à la figure 8. Les étapes relatives à sa mise en œuvre sont les suivantes :

a. constitution d'une grammaire et d'un alphabet. Cette étape permet de générer la population initiale et établit les règles à respecter lors de l'application des opérateurs génétiques;
b. création aléatoire de la première génération. Généralement, le nombre d'individus N de la population initiale reste le même au cours des itérations successives;
c. évaluation de tous les individus (solutions candidates) de la population au moyen d'une fonction d'adéquation (fonction de « fitness »), définie le plus justement possible par le programmeur. Cette évaluation doit refléter la « force » de la solution pour répondre au problème donné. Il est préférable de concevoir une fonction d'adéquation retournant un résultat dans l'intervalle [0;1] afin de faciliter le processus de sélection;
d. sélection des individus pour la reproduction. Plusieurs techniques existent comme le « ranking », les tournois ou encore la roulette; l'objectif étant de favoriser les meilleurs candidats de la population en laissant l'opportunité aux moins bons d'être choisis;
e. la reproduction est un processus prenant en entrée deux individus sélectionnés « parents »; en sortie, elle retourne deux individus « enfants » héritant des caractéristiques de leurs parents. Au cours de ce processus, les opérateurs génétiques de crossover (opérateur binaire) et de mutation (opérateur unaire) interviennent successivement, avec les probabilités respectives p_c et p_m. Dans la plupart des applications, les deux enfants remplacent systématiquement les deux parents à la

génération suivante. Toutefois, certaines variantes proposent d'évaluer les deux enfants et les deux parents à ce stade et de garder les deux meilleurs individus parmi les quatre. Ainsi, la nouvelle génération de solutions candidats est produite, permettant au processus itératif de continuer. L'élitisme est un autre opérateur génétique qui peut être utilisé. Il intervient avant l'étape de sélection et permet de conserver automatiquement une proportion p_e des meilleurs individus à la génération suivante (cf. figure 8); de cette manière, le programmeur s'assure de ne pas perdre la solution optimale si elle est déjà trouvée ou bien une solution proche de celle-ci.

Dans le chapitre 4 consacré à notre algorithme, nous revenons sur les choix effectués lors de ces différentes étapes.

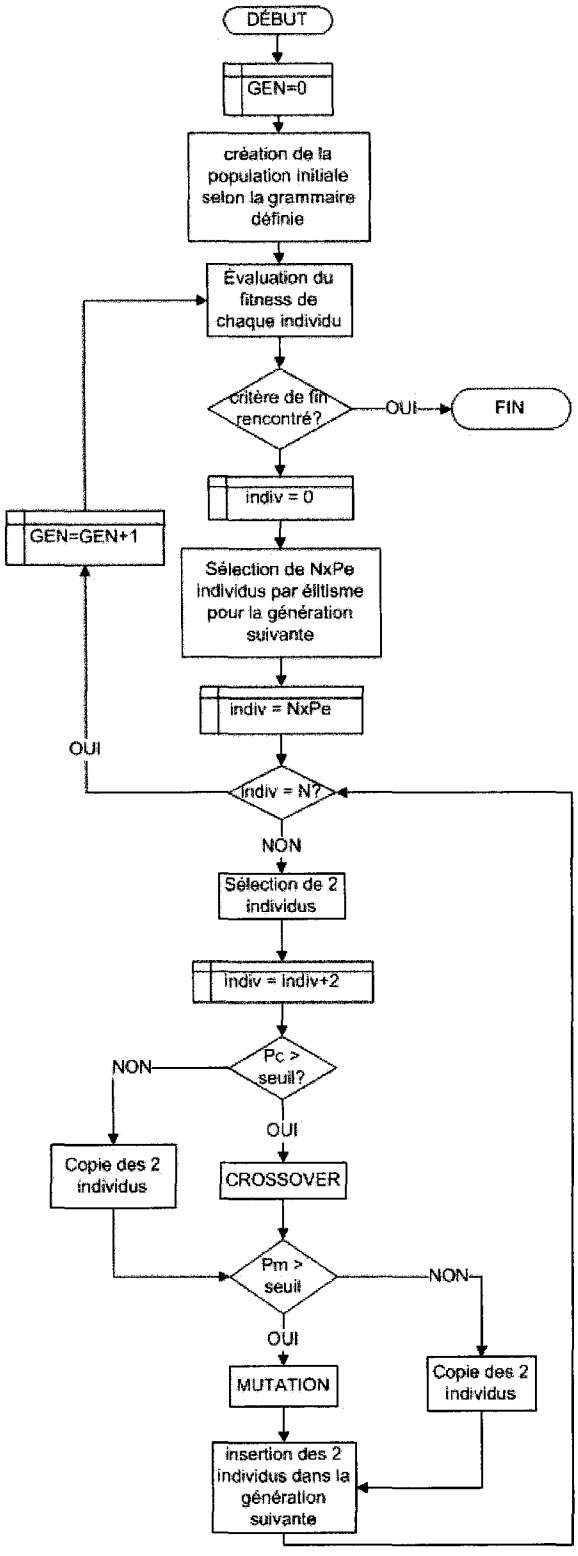


Figure 8 Schéma général de la programmation génétique

CHAPITRE 3

DONNÉES DE L'ÉTUDE

Dans ce chapitre, nous présentons les étapes du processus de préparation des données, depuis leur acquisition jusqu'à leur exploitation par notre algorithme. Il s'agit aussi bien des données de terrain que des données de télédétection.

3.1 Le projet GEOIDE 2000 : présentation générale

Les données utilisées dans cette étude ont été acquises lors d'une campagne de mesures du projet GEOIDE 2000, effectuée au sein de la ferme du Campus Macdonald de l'Université McGill, à Sainte-Anne-de-Bellevue dans l'ouest de l'île de Montréal (cf. figure 23 à l'ANNEXE 2). L'objectif de cette campagne était de mesurer plusieurs variables biophysiques d'un champ de maïs, au cours de sa croissance, afin de les étudier conjointement à des images hyperspectrales acquises par le capteur aéroporté CASI. Pour ce faire, entre avril et septembre, quatre survols ont été effectués dénommés IFC 1 à IFC 4, (« Intensive Field Campaigns » 1 à 4) aux stades clés de développement des plants de maïs, à savoir :

a. IFC 1 : sol nu, avant la pousse;
b. IFC 2 : début du stade végétatif (V1);
c. IFC 3 : début du stade reproductif (R1);
d. IFC 4 : courant du stade reproductif (R).

Pour notre application, nous nous intéressons aux données IFC 3 acquises le 5 août 2000 au début du stade reproductif (R1), caractérisé par l'apparition de fils de soie à

l'extrémité des épis (cf. figure 9). C'est pour cette période que nous disposons du plus grand nombre de données.

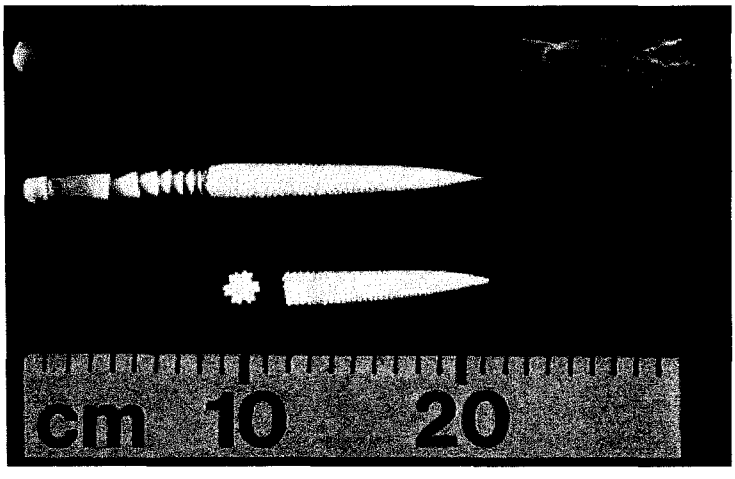


Figure 9 Stade R1 de la croissance du maïs (Université du Dakota du Nord[6])

Au cours de cette campagne, un champ-test a été découpé en parcelles carrées de 20 mètres de côté, chacune traitée par une combinaison d'apport d'azote N_i, i = 0,...,2 et d'herbicide W_j, j = 1,...,4 comme montré à la figure 10; les trois types de traitements d'azote N_0, N_1 et N_2 correspondent respectivement aux applications de a) 60 kg/ha (faible), b) 120 kg/ha (normal) et c) 200 kg/ha (élevé) d'engrais azotés. Concernant les traitements en herbicides W_1, W_2, W_3 et W_4, ils correspondent respectivement à a) aucun traitement, b) traitement des mauvaises herbes feuillues (« broadleaf control »), c) traitement des herbacées (tel le chiendent) et d) traitement total. Au cœur de chaque parcelle, plusieurs mesures ont été effectuées tout au long du développement du maïs. Les principales variables biophysiques mesurées sont données ci-après.

a. mesures de la teneur en azote du sol;

b. analyses de la composition des tissus des plants;

[6] http://www.ext.nodak.edu/extpubs/plantsci/rowcrops/a1173/a1173w.htm

c. mesures de fluorescence;

d. mesures de la verdeur avec le capteur SPAD 502 de Minolta;

e. mesures de la hauteur des plants;

f. mesures de l'humidité du sol;

g. mesures de l'ISF (LAI);

h. mesures de l'azote par un capteur spécifique;

i. mesures spectrales au sol de la canopée;

j. observations de l'infestation par les mauvaises herbes;

k. mesures des récoltes par parcelle.

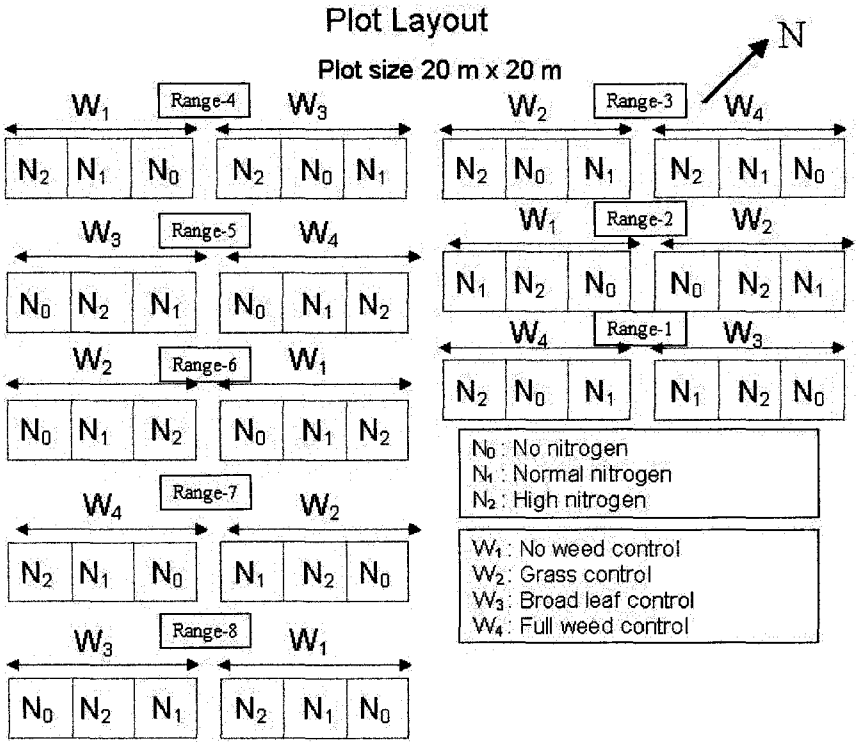


Figure 10 Traitements effectués sur le champ de la ferme du campus Macdonald

Ces différents relevés sont effectués une ou plusieurs fois au cours du développement du maïs, le plus souvent au moment du survol avec le capteur CASI. Dans la section suivante, nous discutons des caractéristiques des données disponibles, utilisées dans ce projet.

3.2 Description de l'approche et détermination des données utiles

L'objectif final de cette étude est de trouver une méthode pour quantifier, à partir d'une image de télédétection, la variabilité en azote dans la canopée d'un champ d'une grande culture. La méthode évolutive brièvement décrite à la section 1.4 nécessite l'emploi de données de terrain pour entraîner l'algorithme (détaillé au chapitre suivant). Ne possédant pas les analyses d'azote effectuées en laboratoire sur les plants de maïs, nous devons trouver un moyen de retrouver cette information d'une autre façon; pour cela, nous proposons une approche utilisant les mesures SPAD (décrites à la section 3.3.1). Dwyer et al. [59] ont établi une fonction f_N (3-4) permettant de transformer une mesure SPAD (sans unité) effectuée sur une feuille de maïs en sa concentration d'azote (en gramme d'azote / kilogramme de feuilles sèches).

Un raisonnement simple nous permet d'établir la relation (3-1) qui donne l'expression QN de la quantité totale d'azote contenue dans un pixel de l'image hyperspectrale. Il serait intéressant de valider cette expression avec des mesures réelles d'azote, mais en attendant, il est raisonnable de penser qu'elle approxime cette valeur.

$$QN = ISF \times f_N(SPAD) \times \rho s_{feuille} \times Aire_{pixel} \qquad (3\text{-}1)$$

Où :

- QN : quantité d'azote totale des feuilles, sur la surface couverte par un pixel, exprimée en g d'azote.
- ISF : Indice de surface foliaire en m^2 de feuilles / m^2 de surface au sol, mesuré avec le LAI-2000 plant canopy analyser.
- SPAD : mesure effectuée avec le SPAD 502 de Minolta, exprimée en unités SPAD (sans dimension).
- $f_N(SPAD) := 0.708 + 0.00898 \times (SPAD) + 0.0004 \times (SPAD)^2$, exprime la quantité d'azote contenue dans une feuille à partir de la mesure SPAD effectuée sur celle-ci (en g / kg de feuilles sèches).
- $\rho s_{feuille}$: densité surfacique d'une feuille en kg / m^2 de feuilles. Il est possible de se représenter cette variable comme le rapport de la masse m d'une feuille sur sa surface.
- $Aire_{pixel}$: surface au sol couverte par un pixel, en m^2 (résolution).

En effet, les mesures SPAD représentant l'activité chlorophyllienne, elles permettent d'estimer la concentration d'azote contenue dans les feuilles d'un plant, qui sont les principaux acteurs intervenant dans le spectre de réflectance du couvert végétal. En pondérant ces concentrations par l'ISF, par l'aire couverte par un pixel (résolution de l'image CASI utilisée, à savoir $4m^2$) et par une valeur $\rho s_{feuille}$ équivalente à la densité surfacique d'une feuille, nous obtenons alors une estimation pertinente de la quantité totale d'azote contenue dans un pixel de l'image hyperspectrale.

Revenons brièvement sur le facteur $\rho s_{feuille}$ énoncé plus tôt; faute de données, il ne peut être évalué pour cette étude. Cependant, cette valeur peut être considérée comme constante, en supposant que l'épaisseur d'une feuille de maïs varie peu et que sa masse volumique est une caractéristique également stable. En effet, si nous considérons la masse m d'une feuille de maïs rapportée à son volume V comme étant une caractéristique intrinsèque stable (en négligeant les variations induites par les différences

de compositions en protéines et autres pigments) et également l'épaisseur ε comme une valeur constante pour une espèce donnée, nous avons alors :

$$\frac{m}{V} = \text{constante } 1 = \frac{m}{\varepsilon \times \text{Surface de la feuille}} = \frac{\rho s_{\text{feuille}}}{\varepsilon} = \frac{\rho s_{\text{feuille}}}{\text{constante } 2} \qquad (3\text{-}2)$$

Ceci impliquant :

$$\rho s_{\text{feuille}} = \text{constante } 1 \times \text{constante } 2 = \text{constante} \qquad (3\text{-}3)$$

Cette hypothèse est renforcée par le fait que les feuilles intervenant dans les mesures sont toutes rendues au même stade de maturité; les variations des deux paramètres ε et m/V seraient vraisemblablement plus importantes si nous n'étions pas dans cette situation.

Par conséquent, cette approche nous permet d'obtenir une estimation de la quantité totale d'azote contenue dans un pixel de l'image CASI, à une constante près qui est le ρs_{feuille}.

Bien que cette simplification entraîne certaines erreurs, elle nous permet de vérifier notre approche.

3.3 Description des données retenues pour l'étude

Conformément à l'approche proposée et expliquée à la section 3.2, les données disponibles et utiles sont détaillées dans les sous-sections suivantes. Lorsqu'il y a lieu, nous décrivons les traitements et prétraitements effectués sur les données brutes pour les corriger et les rendre exploitables.

3.3.1 Mesures SPAD

3.3.1.1 Description

Les mesures SPAD sont des mesures indirectes de l'activité chlorophyllienne des plantes et donc de leur santé. Cet appareil mesure la transmittance de la feuille, au pic d'absorption par la chlorophylle, à la longueur d'onde λ = 650 nm (rouge) et dans la zone de non-absorption de la chlorophylle à la longueur d'onde λ = 940 nm. Le microprocesseur exécute un calcul au moyen de ces deux transmittances pour rendre en sortie une valeur sans dimension (unité SPAD) comprise dans l'intervalle [0 ; 99.9]. Cette valeur est corrélée à 95% aux mesures destructives de chlorophylle effectuées en laboratoire [60]; plus elle est élevée, plus la plante est en bonne santé et inversement.

3.3.1.2 Traitements

Dwyer et al. [59] ont déterminé deux relations non linéaires entre les mesures SPAD et la concentration en azote d'une feuille de maïs, dépendamment que l'on considère les mesures avant ou après le stade reproductif R1. Comme énoncé à la section 3.1, nous travaillons avec les données du vol IFC 3, effectué au début de la phase R1 et pour cette période, les auteurs proposent la relation (3-4).

$$C_{azote} = 0.708 + 0.00898 \times (SPAD) + 0.0004 \times (SPAD)^2 = f_N(SPAD) \qquad (3\text{-}4)$$

Dans cette équation, la concentration en azote C_{azote} est exprimée en g d'azote / kg de matière sèche et la valeur SPAD n'a pas de dimension. Il est à noter que ces relations sont largement reprises comme références par d'autres auteurs [61, 62] utilisant des mesures SPAD.

Nous disposons d'un ensemble de 192 mesures SPAD acquises le 8 août 2000. Au sein des 48 parcelles du champ, quatre drapeaux sont disposés autour de chacun desquels (dans un rayon de 1 mètre environ), cinq relevés SPAD sont effectués sur les feuilles au sommet des plants, sur la 3ème feuille. Étant données les résolutions des images du CASI (1 mètre et 2 mètres pour les deux modes d'acquisition), seule la position centrale du drapeau est enregistrée par GPS, ce qui nous contraint à considérer la valeur moyenne de ces cinq mesures pour notre travail. D'une façon générale, la variabilité des mesures SPAD autour de chaque drapeau est assez faible pour que cette approche soit statistiquement juste; en effet, pour chaque parcelle, le coefficient de variation CV (mesure de dispersion relative) défini comme étant le rapport entre l'écart-type et la moyenne d'une série d'observation X (i.e. $CV(X) = \sigma(X)/\mu(X)$) oscille autour de 10%, ce qui signifie que la moyenne est une « bonne » à « très bonne » estimation des données[7].

Les mesures SPAD ont donc été traduites en concentrations d'azote des feuilles de maïs. L'autre variable biophysique nécessaire pour satisfaire la relation (3-1) est l'ISF, et elle est décrite dans la section suivante.

3.3.2 Mesures de l'ISF

Nous disposons d'un ensemble de 88 mesures d'ISF acquises le 11 août 2000; ces mesures ont été effectuées au moyen du LAI-2000 Plant Canopy Analyzer (LI-COR, Lincoln, NE) autour de quelques-uns uns des drapeaux mentionnés dans la section précédente. Les mesures effectuées au-dessus et en dessous de la canopée sont utilisées pour déterminer l'énergie lumineuse interceptée par le feuillage, et ce pour cinq angles spécifiques; à partir de ces valeurs, l'ISF est calculé en utilisant un modèle de transfert radiatif pour les couverts végétaux.

---

[7] http://www.statcan.ca/cgi-bin/imdb/p2SV_f.pl?Function=getSurvey&SDDS=4706&lang=fr&db=IMDB&dbg=f&adm=8&dis=2

Disposant de 88 points de mesure d'ISF, nous sommes donc limités à cette quantité de valeurs de référence pour entraîner et valider notre algorithme. Il nous reste donc à présenter les données CASI utilisées pour l'étude et c'est le sujet de la section suivante.

3.3.3 Données CASI

3.3.3.1 Description

Les images hyperspectrales de l'étude ont été acquises par le capteur aéroporté CASI selon les deux modes décrits à la toute fin de l'ANNEXE 1. Comme énoncé plus tôt dans ce chapitre, les données hyperspectrales de notre travail sont celles correspondant au vol IFC 3 effectué le 5 août 2000, puisque c'est autour de cette période du développement du maïs que nous disposons du plus grand nombre de mesures au sol.

Des deux modes d'acquisition décrits dans la même annexe, nous utilisons celui dont les caractéristiques sont les suivantes :

a. 72 bandes spectrales contiguës de largeur 7.47 nm allant de 409 nm à 947 nm;
b. résolution spatiale de 2 m × 2 m, soit une aire totale de 4 m² pour un pixel.

Cet ensemble de données est celui qui comporte le plus d'information (72 caractéristiques par pixel au lieu de 7 pour l'autre mode, tel qu'indiqué par le vecteur de caractéristique donné en (A-3)) et c'est donc celui qui est le plus susceptible de satisfaire notre recherche d'informations discriminantes dans le spectre de réflectance pour caractériser la variabilité en azote.

3.3.3.2 Prétraitements

Cette partie de la préparation des données a été effectuée en amont de mon travail par l'équipe de recherche du Dr. John Miller du Centre for Research in Earth and Space Science (CRESS) de l'Université de York de Toronto en Ontario. Au cours de cette étape, les images ont subi les trois types de corrections nécessaires à leur utilisation pour une analyse quantitative, à savoir :

a. une correction radiométrique : les images sont d'abord transformées en données de radiance relative au capteur grâce à l'utilisation de coefficients de calibrage déterminés au sein du laboratoire;
b. une correction atmosphérique : par la suite, l'utilisation d'un modèle de correction atmosphérique est utilisé pour transformer les données relatives de radiance en données absolues de réflectance. Des matériaux de référence, présents sur les images, dont les réflectances sont parfaitement connues (habituellement béton ou asphalte) sont couramment utilisés à cette étape de correction. Dans notre cas, des marqueurs en TYVEK blanc, placés aux extrémités du champ, ont été utilisés comme référence pour cette étape de correction;
c. une correction géométrique : enfin, les mouvements parasites de l'avion au cours du vol étant enregistrés par des systèmes embarqués, il est possible d'en corriger les déformations résultantes. Des relevés effectués au GPS au sol permettent par la suite le géoréférencement des images.

Ces étapes ne faisant pas partie de notre travail, nous n'irons pas plus loin dans leur description.

3.3.3.3 Traitements

L'étape de prétraitement est indispensable afin de rendre les données exploitables pour une analyse numérique. Toutefois, ces rectifications ne sont pas parfaites, tant sur le point radiométrique que géométrique. De plus, lorsque les points d'échantillonnage au sol (ISF et SPAD) ont été référencés par GPS, de nouvelles erreurs de positionnement (cf. figure 11) viennent s'ajouter aux précédentes erreurs énoncées. Avant d'utiliser les images, nous leur appliquons par conséquent un filtre moyennant 3×3 afin de réduire l'influence de ces erreurs.

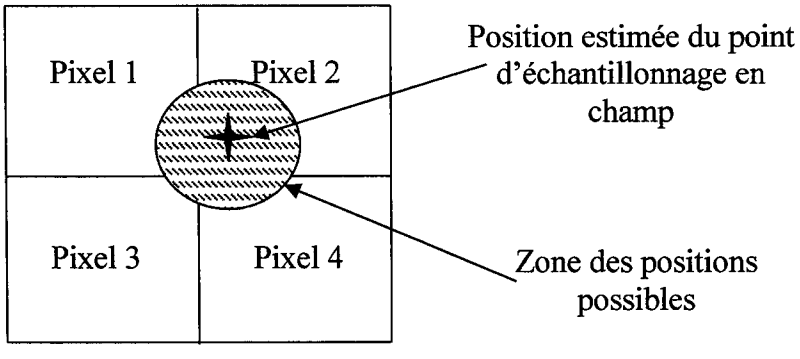


Figure 11 Erreur de positionnement- Exemple sur 4 pixels voisins

Sur l'exemple de la figure 11, une erreur de positionnement est possible entre les pixels 1 à 4 et donc, d'une façon générale cette erreur peut impliquer chacun des 9 pixels adjacents. De plus, comme il y a en réalité cinq mesures SPAD effectuées dans un rayon de 1 mètre, la valeur des pixels voisins doit intervenir pour représenter correctement cette moyenne des valeurs SPAD.

Il n'y aura pas de problème d'effets de bords (influence de la terre apparente des allées entre les parcelles) lors du moyennage puisque les points d'échantillonnage se situent au cœur des parcelles ayant subi les mêmes traitements en azote; les pixels adjacents sont par conséquent à l'intérieur des parcelles. L'application d'un filtre moyennant permet également de réduire le bruit (de mesure entre autres) de ces images.

Le dernier traitement effectué est la suppression des deux dernières bandes spectrales (numéro # 71 et # 72) puisque ces bandes sont corrompues et ne véhiculent donc aucune information pertinente pour notre étude. Nous nous retrouvons par conséquent avec un ensemble de 70 bandes spectrales correspondant aux longueurs d'onde allant du violet 409 nm au proche infrarouge 932 nm.

3.4 Création d'une base de données géoréférencée

3.4.1 Implantation des données

Pour extraire les données utiles, nous devons commencer par créer une base de données géographique au sein d'un logiciel de SIG. Cette étape est nécessaire pour recueillir les valeurs des 70 bandes de l'image hyperspectrale aux coordonnées géographiques des données de terrain SPAD et d'ISF.

Cette base de données est constituée au sein de PCI GEOMATICA version 9.1.6. Les deux types de données dont nous disposons sont a) des données vectorielles (données de terrain) et b) des données matricielles (bandes de l'image hyperspectrale).

Les images prétraitées fournies par le CRESS sont en format « .lan », caractéristique des fichiers images traités par un logiciel de la compagnie ERDAS. Le logiciel PCI GEOMATICA utilisé est capable de lire ces données sans aucune modification. De plus

ces données étant d'ores et déjà référencées géographiquement, il ne nous reste qu'à les convertir dans le format « .pix » propre à PCI pour en faciliter l'accès, l'exploitation et l'intégration dans une base de données; par la suite, nous appliquons un filtre moyennant 3x3 pour les raisons évoquées à la section 3.3.3.3.

Les données vectorielles doivent être implantées dans PCI dans le format GAV[8] (Generic ASCII Vector format). Pour ce faire, il faut constituer un « fichier schéma » (« schema file ») avec l'extension « .gav » qui contient les informations sur le contenu et le descriptif des données vectorielles, ainsi qu'un « fichier de données » (« data file ») avec l'extension « .txt » contenant les valeurs des données, organisées comme décrites dans le premier fichier. Dans notre cas, les attributs des données de terrain échantillonnées sont quantitatifs (mesures SPAD transformées et ISF); ils doivent être précédés dans le fichier de données par les coordonnées géographiques de latitude, longitude et altitude dans cet ordre. Pour de plus amples informations sur ce format vectoriel de données, le lecteur pourra se référer au site placé à la note de pied de page[8].

Une fois la base de données géographique créée, toutes les manipulations souhaitées sont réalisables à partir de l'environnement FOCUS de PCI GEOMATICA. Cela inclut l'application d'algorithmes, déjà existants ou bien implémentés par l'utilisateur, sur les éléments de la base. Cela permet également la visualisation des données sous diverses formes, utile pour l'extraction de premières informations qualitatives, purement visuelles, comme montré aux figures 24 et 25 de l'ANNEXE 3.

3.4.2 Sélection et extraction des données

Une fois la base de données réalisée dans l'environnement du logiciel de SIG PCI GEOMATICA, il ne nous reste qu'à échantillonner les bandes spectrales aux

[8] http://www.pcigeomatics.com/cgi-bin/pcihlp/GAV

coordonnées des 88 mesures d'ISF et SPAD, données que nous allons utiliser par la suite au sein de notre environnement de programmation VISUAL C++.

L'algorithme VSAMPLE de PCI permet d'échantillonner les pixels des données matricielles (bandes spectrales dans notre cas) aux points définis par des données vectorielles. Cet algorithme nous permet donc de recueillir l'information des pixels des bandes spectrales en chacun des points de mesures ISF et SPAD.

Les signatures spectrales des pixels échantillonnés possèdent une variabilité importante, particulièrement dans la zone du proche infrarouge et moindrement dans le vert (cf. figure 12). Nous reconnaissons toutefois la forme générale de la réponse spectrale des végétaux présentée à la figure 2 de la section 1.3.

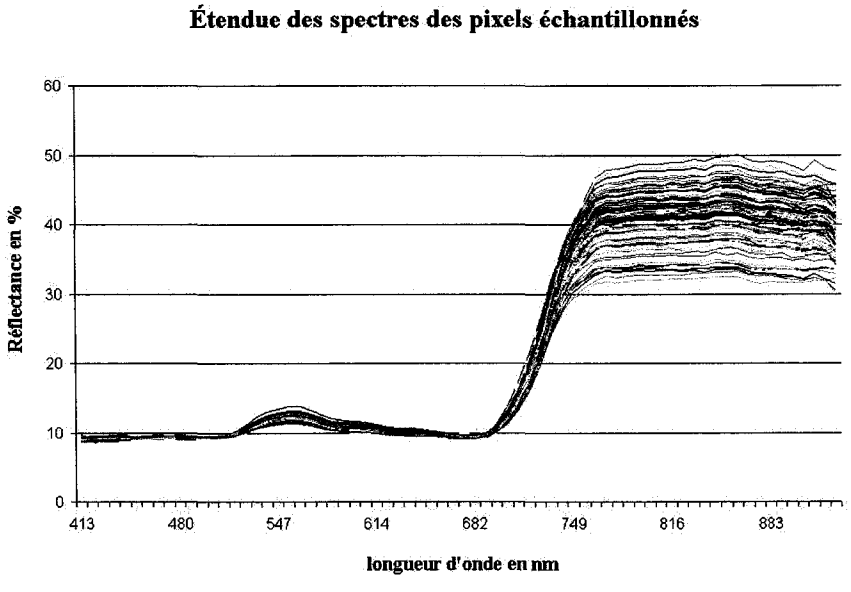


Figure 12 Variabilité spectrale des 88 pixels échantillonnés.

3.5 Limites des données

La première limite évidente du jeu de données utilisé provient du fait qu'elles ont été acquises à des dates différentes. Bien que les dates soient assez rapprochées (réparties sur une semaine), cela apporte une variabilité parasite qui vient bruiter nos données de travail.

Une autre sorte de bruit est induite par les traitements en herbicides W_i variables sur l'ensemble du champ-test. Bien que l'objectif de notre étude soit d'être capable d'isoler la variabilité en azote des autres paramètres biophysiques du champ, cette variabilité en herbicides (et donc en présence de mauvaises herbes) n'est pas « naturelle » et exagérée comparativement au cas d'un champ réel. De plus, la présence de mauvaises herbes implique une augmentation de l'ISF ainsi que de la quantité d'azote, sans que cela ne soit pris en compte par les mesures effectuées au sol. Cela vient par conséquent ajouter un bruit supplémentaire à nos données.

Pour finir, les références en azote prises au sol (« ground truth ») sont estimées à partir d'un modèle empirique [59] appliqué à des mesures SPAD. Accepté par la communauté, l'instrument SPAD 502 de Minolta est surtout apprécié pour ses qualités pratiques; il est en effet facile à utiliser et non destructif pour une évaluation rapide de l'activité chlorophyllienne des plantes. Cependant, plusieurs auteurs indiquent des variabilités dans les mesures, dépendant de l'irradiance [63] (10% à 15%), du moment de la journée auquel est effectuée la mesure ainsi que de la teneur en eau des feuilles [64]. Ces estimations d'azote comportent par conséquent des erreurs et donc un bruit supplémentaire.

Pour une analyse plus précise, il serait préférable d'utiliser des mesures d'azote effectuées en laboratoire sur des feuilles prélevées le jour de l'acquisition de l'image hyperspectrale.

3.6 Conclusion

En agronomie, les données de terrain sont coûteuses et difficiles à obtenir en grande quantité. Celles dont nous disposons possèdent des limites (section 3.5) vis-à-vis de l'objectif fixé mais elles restent pertinentes compte tenu des traitements et transformations qu'elles ont subies. La phase de « nettoyage » des données présentée dans ce chapitre est indispensable comme préalable à l'analyse qui va suivre; cette étape de préparation nécessite des connaissances contextuelles sur l'ensemble de données brutes, primordiales pour en extraire la partie significativement exploitable.

CHAPITRE 4

ALGORITHME

Dans ce chapitre, la diversité n'est pas mesurée et elle fait donc référence à une notion intuitive. La répétition importante de « sous-arbres » identiques au sein d'une population caractérise une perte de diversité. Cette notion est évaluée de façon empirique par simple observation visuelle de la syntaxe des individus, par échantillonnage ou par l'étude de petites populations.

4.1 Objectif et stratégie

L'objectif de notre algorithme est de trouver de nouveaux indices de végétation (IV) qui décrivent la variabilité en azote au sein de la canopée d'un champ d'une grande culture; un IV étant une combinaison arithmétique de bandes spectrales, c'est dans cet espace que notre algorithme va chercher une solution.

Un IV est une combinaison arithmétique de bandes spectrales et par conséquent, il peut être représenté par une image en tons de gris. Un moyen d'évaluer si un IV décrit la variabilité en azote est de mesurer la force de la corrélation entre les mesures d'azote effectuées en champ et les valeurs des pixels de cet indice, aux points d'échantillonnage. Dire qu'il existe une forte corrélation entre des mesures d'azote et un IV est alors équivalent à considérer qu'il existe un modèle de régression qui relie les données avec un coefficient de détermination R^2 élevé, comme montré par la relation (4-1) ci-dessous.

$$QN(x,y) \approx f_{reg}\left(IV(x,y)\right), \text{avec } R^2 \text{ élevé} \qquad (4\text{-}1)$$

Dans cette relation (4-1), QN(x,y) représente la quantité d'azote évaluée au pixel ayant les coordonnées géographiques (x,y); IV(x,y) est la valeur du ton de gris de l'image représentant l'indice de végétation (IV) considéré, aux coordonnées (x,y), et f_{reg} est la fonction mathématique du modèle de régression qui relie les deux jeux de données.

Pour un type de régression donné (linéaire, exponentiel...), c'est le coefficient de détermination R^2 que nous allons considérer pour juger de l'aptitude d'un IV à décrire la variabilité en azote au sein de notre champ.

Nous cherchons donc une combinaison de bandes spectrales qui soit reliée à un sous-ensemble des données d'azote du champ (base d'apprentissage); par la suite, nous évaluons les performances du modèle de régression trouvé sur le reste des données d'azote (base de validation). Les résultats sont présentés au chapitre 5.

4.2 Structure

La programmation de l'algorithme est faite en langage C dans l'environnement de VISUAL C++. Il est à noter que le code de notre algorithme est réalisé de façon à être entièrement indépendant des données utilisées.

La structure générale de notre algorithme est celle de la programmation génétique (PG) présentée à la figure 8. Il faut noter qu'une application utilisant les algorithmes évolutionnaires requiert une bonne compréhension du problème afin d'adapter et même de créer des opérateurs spécifiques pertinents. Nous allons détailler et expliquer, dans les prochaines sous-sections, les choix effectués pour chacun des éléments-clés et chacune des étapes qui font la spécificité de cette technique évolutive.

4.2.1 « Grammaire et alphabet »

La première étape en PG est de définir l'espace dans lequel se situe la solution recherchée. Pour cela, il faut définir des opérateurs sur nos données (« alphabet ») et des règles d'utilisation (« grammaire ») pour créer notre population initiale et la faire évoluer par la suite. Dans notre cas, nous cherchons une combinaison de bandes spectrales (IV) et nous allons donc utiliser les opérateurs arithmétiques simples (4-2) et les 70 bandes spectrales comme données (4-3).

$$\{+, -, \times, \div\} \tag{4-2}$$

$$\{b_i\}_{i\in[1;70]} \tag{4-3}$$

Les individus qui composent les populations sont représentés sous forme d'arbres binaires, comme illustré dans la partie 4.2.2.

Concernant les lois de construction, elles sont simples également; puisqu'il s'agit de faire des opérations entre des bandes spectrales, nous avons les règles suivantes :

a. un opérateur est suivi par :
    - 2 bandes spectrales ou
    - 1 bande spectrale ET 1 opérateur ou
    - 2 opérateurs.
b. une bande spectrale n'est suivie par aucun élément (élément terminal).

Pour mieux comprendre la façon dont ces règles de construction agissent, il faut parler du mode de représentation utilisé pour décrire les individus; c'est ce qui fait l'objet de la section 4.2.2 suivante.

4.2.2 Représentation des individus : les arbres binaires

La représentation utilisée pour définir les individus de nos populations est la structure en arbres binaires. C'est une structure nodale pour laquelle chaque nœud possède une valeur appartenant à l'un des deux ensembles définis par (4-2) et (4-3), et est suivi par deux « nœuds fils ». Cette structure permet de représenter tout type de combinaison arithmétique; par exemple l'indice NDVI classique défini à la section 2.4.1 est décrit par l'arbre binaire de la figure 13; dans cet arbre, les bandes utilisées sont celles du capteur CASI correspondant aux longueurs d'ondes 670 nm (b_{35}) et 800 nm (b_{53}).

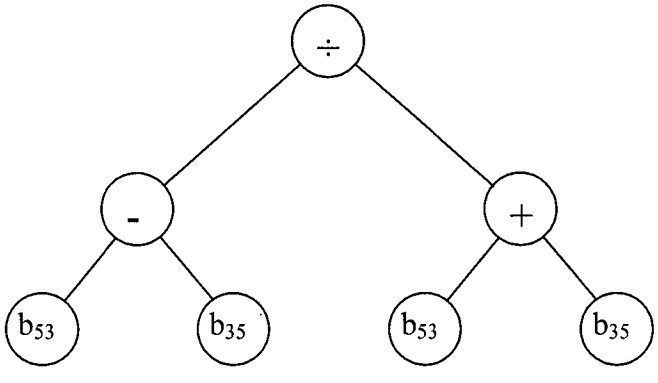


Figure 13 Indice NDVI représenté par un arbre binaire

Il faut toutefois définir une convention pour le sens de lecture de l'arbre tel que représenté à la figure 13; dans notre cas, sous un nœud opérateur, l'élément de gauche

est prioritaire sur celui de droite. Ainsi, il n'y a plus de confusion possible et cet arbre sera traduit par la relation (4-4).

Nous optons pour ce mode de représentation en raison de la simplicité à représenter les solutions de notre espace de recherche.

$$NDVI = \frac{b_{53} - b_{35}}{b_{53} + b_{35}} \tag{4-4}$$

À présent que sont définis les opérateurs, les données et le mode de représentation des individus, nous allons entrer dans la description du déroulement de l'algorithme.

4.3 Fonctionnement de l'algorithme

4.3.1 Population initiale

4.3.1.1 Taille N_p de la population

Comme décrit dans la section 2.6.2, la première étape de la PG est la création aléatoire de la première génération d'individus. La taille N_p (nombre d'individus) de la première population est d'abord déterminée selon le type de problème (caractéristiques de l'espace de recherche, complexité du problème...) puis, elle peut être ajustée de façon empirique au cours des expériences, notamment en raison du temps de calcul. Pour notre problème, nous gardons le paramètre N_p constant au cours des itérations d'une simulation; d'une façon générale, nous prenons ce paramètre dans l'intervalle donné en (4-5).

$$N_p \in [100 ; 2000] \quad (4\text{-}5)$$

Une valeur trop faible du paramètre N_p limite le potentiel de diversité et donc réduit les chances de trouver la solution optimale par restriction trop forte de l'espace de recherche. Une valeur trop élevée entraîne premièrement des temps de calcul excessifs et de surcroît, risque de ralentir considérablement la vitesse de convergence.

Une discussion sur ce paramètre sera faite lors de la présentation des simulations et des résultats au chapitre 5.

4.3.1.2 « Profondeur maximale » de la première génération

En programmation génétique (PG), un problème largement documenté [65-68] souvent résolu au cas par cas est l'explosion de la taille des individus au cours des itérations. Ce phénomène connu sous le nom de « code bloat » provoque l'envahissement de l'espace mémoire du PC sur lequel roule la simulation et peut entraîner la destruction du processus.

Un autre problème fréquemment rencontré au cours de l'entraînement d'algorithmes d'apprentissage, est le surapprentissage désigné par le mot anglais « overfitting ». Ce phénomène se traduit par une perte en capacité de généralisation des solutions. En PG, le surapprentissage se manifeste par une augmentation rapide de la taille des individus accompagnée d'un très faible gain des valeurs de fitness (cf. section 4.3.2). A ce stade, l'augmentation de la taille des individus peut également être préjudiciable.

Pour notre application, nous privilégions la recherche d'une solution courte en terme du nombre d'éléments. De plus, étant donnés les problèmes énoncés ci-dessus, nous décidons de créer un paramètre afin de limiter la profondeur des individus de la première

génération; la profondeur étant définie par le nombre de niveaux de nœuds d'un individu (à la figure 13, la profondeur est égale à 3). De ce fait, nous privilégions une première population constituée de nombreux individus courts, dans le but de sélectionner rapidement de petits individus constitués d'information discriminante. Nous faisons confiance aux opérateurs génétiques pour composer des individus plus complexes par la suite.

Ainsi, lors de la création de la population initiale, la profondeur maximale la plus fréquemment choisie lors des simulations est 3 (comme à la figure 13).

Une autre précaution prise lors de la création de cette population initiale est l'élimination des individus « monobandes », c'est à dire composés d'une seule bande (pas d'opérateur). Ces individus n'apportent rien de plus qu'une simple mutation lors des croisements. Ils affectent indirectement la diversité de cette première génération en prenant la place d'individus plus intéressants.

4.3.2 Fonction d'adéquation ou de « fitness »

La fonction d'adéquation a pour but d'évaluer la justesse avec laquelle un individu répond au problème. Pour chaque génération, tout individu possède une seule valeur de fitness qui va être utilisée à l'étape de sélection (cf. section 4.3.3). Le choix de cette fonction doit représenter le plus justement possible le critère que l'on souhaite satisfaire. Pour notre application, nous cherchons à trouver la combinaison de bandes spectrales la plus fortement corrélée à la quantité d'azote présente dans le champ; cela signifie que nous cherchons à maximiser la valeur du coefficient de détermination R^2 associé au modèle de régression f_{reg} (cf. la relation (4-1)) qui relie un individu IV aux mesures d'azote QN_k (cf. (4-6) ci-dessous) de notre base d'entraînement.

$$IV = \begin{bmatrix} IV_1 \\ IV_2 \\ \cdots \\ IV_L \end{bmatrix} \text{ et } QN_k = \begin{bmatrix} QN_1 \\ QN_2 \\ \cdots \\ QN_L \end{bmatrix} = y_k \qquad (4\text{-}6)$$

$(IV_k)_{1\le k\le L}$ est le vecteur des valeurs de l'indice IV aux points de mesures d'azote QN_k et L est la taille de la base d'entraînement. Nous discutons des concepts de base d'entraînement et de base de validation dans le chapitre 5.

Pour cela, nous considérons quatre types de régressions classiques pour chacune desquelles nous allons conduire des expériences distinctes. Ces régressions sont :

a. linéaire;
b. exponentielle;
c. logarithmique;
d. puissance.

Cette valeur du coefficient de détermination R^2 intervient dans le processus de sélection et c'est le sujet de la section suivante. Notre valeur de « fitness » possède l'avantage d'être normalisée puisque R^2 est compris dans l'intervalle [0 ; 1], la valeur 1 étant l'objectif à atteindre.

4.3.3 Sélection

La sélection est une étape critique dans la recherche de notre solution. Elle doit idéalement favoriser les meilleurs individus de la population en laissant leur chance aux

moins bons, pour éviter de perdre de la diversité. Nous testons plusieurs types de sélection dont les principaux utilisés sont détaillés ci-après.

4.3.3.1 Roulette simple

Ce type de sélection choisit un individu k avec la probabilité P(k) donnée à la relation (4-7), où R_k est le coefficient de Pearson associé au modèle de régression discuté plus haut dans ce chapitre. Ici, nous préférons utiliser la valeur absolue du coefficient de Pearson ($|r_k| = |R_k|$) au coefficient de détermination R^2 car ce dernier varie avec plus d'amplitude que $|r_k|$ sur la plage des valeurs qui nous concerne; il laisse par conséquent plus de chance à un individu légèrement moins bon qu'un autre en « écrasant » l'écart entre eux.

$$P(k) = \frac{|R_k|}{\sum_{i=1}^{i=N} |R_i|}, \quad |R_k| = +\sqrt{R_k^2} \tag{4-7}$$

Lors de l'utilisation de cette méthode de sélection, nous prenons comme précaution de ne pas sélectionner deux individus identiques pour se reproduire, évitant ainsi une forme de « consanguinité » qui n'apporte rien de neuf à notre future génération; c'est fréquemment lorsque deux solutions identiques se croisent qu'apparaissent des éléments longs et surentraînés sur les données, ce qui peut également provoquer une convergence hâtive vers un maximum local.

Nous constatons que cette méthode de sélection est trop élitiste et réduit rapidement la diversité au sein des populations, c'est pourquoi elle ne sera pas utilisée seule.

4.3.3.2 Tournoi à 4

Par hasard au 1er tour et roulette au 2ème tour

Cette approche permet de conserver de la diversité au sein de la population. De ce point de vue, elle est intéressante d'autant que nous prenons garde à ne pas sélectionner 2 individus identiques au 1er tour ni au 2ème tour pour la reproduction. Toutefois, nous observons une explosion de la taille des individus.

Par roulette au 1er tour et minimisation de la longueur au 2ème tour

Ce mode de sélection choisit d'abord 4 solutions différentes parmi toute la population selon le mode de sélection par roulette (décrit au point précédent); par la suite, nous en sélectionnons 2 différents parmi ces 4 pour la reproduction, avec pour chaque individu k, la probabilité P(k) (cf. relation (4-8)) d'être sélectionné.

Dans cette relation (4-8), L_i représente la longueur de l'individu i. Ainsi, on privilégie un individu court lors de ce 2ème tour du tournoi. Intuitivement, nous comprenons qu'une solution qui généralise bien un problème doit être la plus simple possible. Une solution trop complexe (donc trop longue) avec un coefficient R^2 élevé aura tendance à « coller » trop fortement aux données d'entraînement. Prendre ce paramètre en compte nous donne l'opportunité d'éviter de converger trop rapidement vers des maxima locaux.

$$P(k) = \frac{\sum_{i=1}^{N} L_i}{L_k} \times \frac{1}{\sum_{j=1}^{N} \left(\frac{\sum_{i=1}^{N} L_i}{L_j} \right)} = \frac{1}{L_k \times \left(\sum_{j=1}^{N} \frac{1}{L_j} \right)} \qquad (4\text{-}8)$$

En observant le comportement des simulations, le mode de sélection au 2ème tour s'avère trop radical quant à la réduction de la taille des individus; en effet, nous arrivons fréquemment à une population finale dans laquelle la moyenne des tailles de tous les individus est légèrement supérieure à 3 (3 étant la taille minimum d'un individu, en nombre de nœuds, puisqu'il n'y a pas de solutions « monobandes » au sein des populations, cf. 4.3.1.2). De plus, une autre tendance était la perte d'individus ayant un R^2 élevé et une longueur légèrement supérieure aux autres candidats du tournoi. Pour remédier à ce problème, nous cherchons à « adoucir » cette sélection sur la longueur et pour ce faire, nous proposons la probabilité P(k) décrite par la relation (4-9).

$$P(k) = \frac{\left|R_k\right| \Big/ \left[0.4 + \log(1 + L_k)\right]}{\sum_{i=1}^{N} \left\{ \left|R_i\right| \Big/ \left[0.4 + \log(1 + L_i)\right] \right\}}, \quad 0 \le R_i \le 1 \text{ et } L_i \ge 3 \qquad (4\text{-}9)$$

De cette manière, nous gardons le poids du coefficient de corrélation R_k prépondérant; en effet, comme $L_i \geq 3$, la valeur $\left[0.4 + \log(1 + L_k)\right] \geq 1{,}0020$ et croît de façon logarithmique lorsque L_i augmente. De cette façon, nous avantageons les individus bons et courts. C'est ce mode de sélection que nous utilisons pour nos simulations.

À présent que nous pouvons sélectionner nos individus d'après nos besoins, nous allons décrire les opérateurs qui servent à créer les populations à partir de ces éléments choisis.

4.3.4 Opérateurs génétiques

4.3.4.1 Élitisme

Comme montré à la figure 8, l'élitisme intervient indépendamment du processus de sélection décrit à la section 4.3.3. Cet opérateur sélectionne les E meilleurs individus de la génération *n* et les insère automatiquement dans la génération *n+1*; cela nous assure de ne pas perdre la meilleure solution si elle a déjà été trouvée avant la fin de la simulation. Ce nombre E est donné par la relation (4-10).

$$E = p_e \times N_p \tag{4-10}$$

Dans cette relation (4-10), p_e est le pourcentage d'élitisme et N_p est la taille de la population. À cette étape, il faut faire attention à ce que la quantité $N_p \times (1 - p_e)$ soit paire puisqu'il s'agit du nombre d'individus restant à sélectionner pour la reproduction.

Il existe d'autres formes d'élitisme mais nous nous contentons de celle-ci qui est la plus classique. Une autre forme classique est l'élitisme faible qui remplace le moins bon élément de la génération *n+1* par le meilleur de *n* si ce dernier est meilleur que le meilleur de *n+1*. Cette méthode n'est pas testée mais elle est suggérée dans le chapitre sur les recommandations.

4.3.4.2 « Crossover »

Lorsque deux individus ont été sélectionnés pour la reproduction, il y a alors une probabilité p_c pour qu'il y ait l'application de la fonction de Crossover entre ceux-ci. Lorsque ce croisement a lieu, un endroit de section est choisi chez les deux individus et un échange des sous-arbres inférieurs est alors effectué, comme illustré à la figure 14. Ainsi, les deux enfants créés possèdent des caractéristiques des deux parents.

Pour cet opérateur, il existe des variantes; ces formes de crossover plus complexes ne sont pas jugées pertinentes pour notre application dans la mesure où elles alourdiraient le code et par la même occasion augmenteraient le temps de calcul des simulations.

Généralement, la probabilité d'occurrence du crossover est élevée ($p_c > 90\%$) puisque cet opérateur est la partie centrale du processus de génération des nouvelles populations.

4.3.4.3 Mutation

Contrairement au crossover, la probabilité d'occurrence de la mutation est généralement faible ($p_m < 10\%$). En intervenant, la mutation apporte de la diversité en introduisant une variabilité d'information extérieure à la population. Elle peut également permettre de donner du sens à un sous-arbre en détruisant un intron comme montré à la figure 15.

Pour notre application, nous utilisons un opérateur de mutation simple; nous commençons par choisir un nœud au hasard et nous remplaçons le sous-arbre dont il est le sommet par un nouveau sous-arbre généré aléatoirement.

Il existe d'autres formes de mutations mais l'importance de cet opérateur n'étant pas de premier ordre, et pour ne pas surcharger le code, nous n'irons pas plus loin sur ce point.

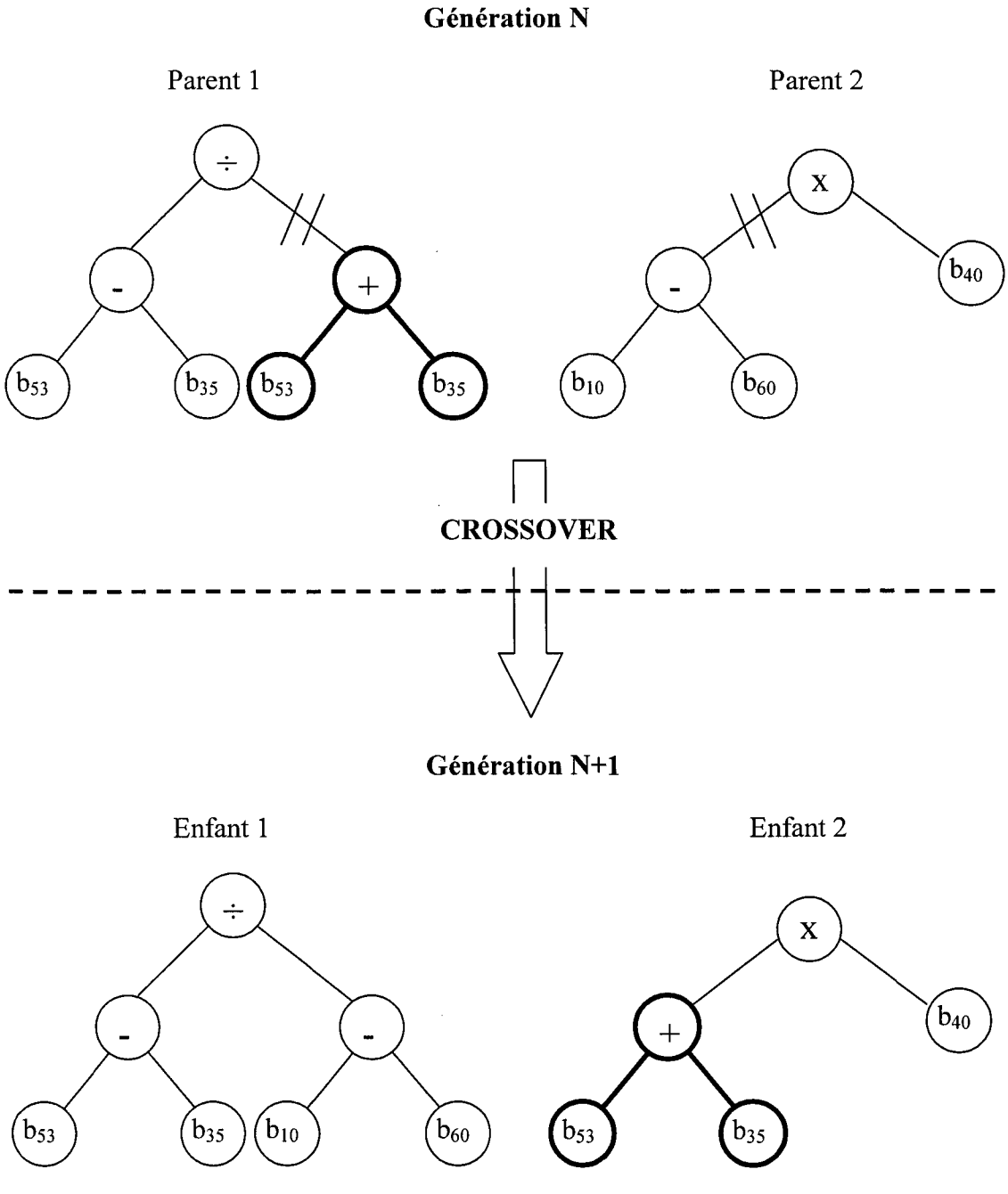


Figure 14 Crossover entre 2 indices de végétation représentés par des arbres.

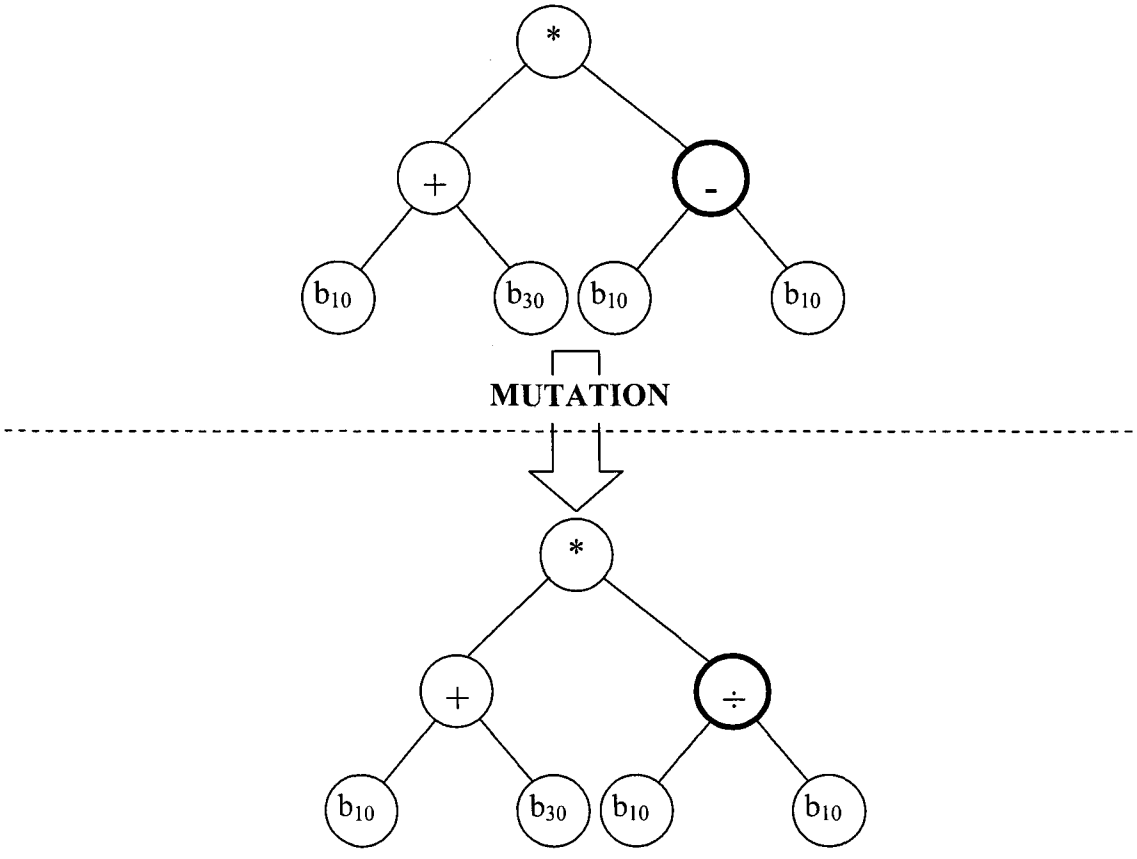


Figure 15 Mutation ponctuelle sur un indice de végétation

4.4 Mesures particulières : optimisation de la recherche

4.4.1 Diversité

Comme il a été mentionné plus tôt dans ce chapitre, un des enjeux pour que la recherche d'une solution soit efficace est le maintien de la diversité au sein de chaque population; cette diversité est essentielle pour permettre à la recherche parallèle d'être effectuée et d'éviter la convergence hâtive vers un maximum local. À ce sujet, nous avons pris plusieurs précautions comme la suppression des individus monobandes ou encore

l'interdiction de faire se reproduire des individus identiques (i.e. pas de « consanguinité »). Pour œuvrer dans ce sens, à chaque nouvelle génération, nous éliminons tous les doublons et les remplaçons par un nouvel individu créé avec les paramètres utilisés pour la population initiale. L'idéal serait d'introduire une mesure de diversité dans notre algorithme afin d'avoir un suivi de ce paramètre au fil des générations; ceci est suggéré dans le dernier chapitre sur les recommandations; toutefois, c'est au-delà de cette recherche.

4.4.2 Surapprentissage

De même, pour éviter le surapprentissage, nous introduisons un paramètre de taille maximale pour les individus. Après quelques simulations et l'observation de la syntaxe des individus, nous sommes en mesure d'évaluer une valeur maximale du nombre de nœuds $N_{n\_max} \approx 30$. Au-delà de ce seuil, l'amélioration du fitness est synonyme de surapprentissage, caractérisée par une augmentation systématique de l'erreur de généralisation définie comme l'erreur RMSE calculée sur la base de validation; nous cherchons ainsi la solution la plus générale possible.

L'interdiction de faire se reproduire deux individus identiques entre eux permet également de contrer le surapprentissage.

4.4.3 Combattre le « code bloat »

Nous l'avons déjà mentionné, un problème connu de la programmation génétique est le « code bloat »; ce terme correspond au phénomène d'expansion de la taille des individus au cours de l'évolution des populations et ce, sans amélioration significative des valeurs de fitness. La raison principale de ce phénomène est la prolifération exponentielle des

introns (multiplications par 0, sous-arbre ne contribuant pas ou peu au calcul du fitness...) [69].

Lorsque le crossover intervient entre deux individus, la plupart du temps, l'enfant est assez différent de ses parents; en avançant dans les itérations, il est de moins en moins facile de trouver de meilleures solutions et c'est à ce moment que l'évolution favorise le « bloat ». En effet, la probabilité d'avoir un enfant avec un fitness au moins aussi bon que celui de ses parents est plus grande lorsque les individus sont longs (avec seulement une petite partie codant de façon significative) que lorsqu'ils sont courts (pour lesquels la majeure partie est significative) puisque dans ce cas-ci, il y a moins de chance que le crossover ne vienne détruire l'individu. Il est donc possible de conclure que le « bloat » est intimement lié au crossover. Pour éviter ce phénomène, il existe principalement quatre moyens[9] :

a. Koza propose des fonctions générées automatiquement qui encapsulent des morceaux d'individus afin de les protéger lors du crossover;
b. le grossissement des populations peut permettre de converger vers une bonne solution avant le bloat;
c. donner une pénalité aux individus trop longs ou bien à ceux ayant de grosses parties non-codant : la pression parcimonieuse [68]. Le double tournoi (fitness + taille) ou bien l'optimisation multicritère sur le fitness, taille et diversité apporte aussi une réponse au problème;
d. enfin, en donnant plus d'importance à la mutation et moins au crossover mais cela comporte le risque de dénaturer la spécificité de la recherche.

Toutes ces méthodes (et bien d'autres non mentionnées) ne règlent pas le problème du « code bloat »; dans la plupart des cas, elle le restreint mais il n'est pas encore possible d'empêcher ce problème sans contraindre excessivement la recherche de la solution.

[9] http://www.meteo.uni-bonn.de/mitarbeiter/venema/essays/2004/genetic_programming_and_bloat.html

Pour notre application, nous utilisons la pénalisation des individus trop longs, par la méthode de sélection et la suppression de ceux dépassant une valeur-seuil, comme mentionné à la section 4.4.2.

4.5 Résumé des paramètres de simulation

Nous présentons dans le tableau I, l'ensemble des paramètres de simulation ainsi que leur intervalle d'utilisation, choisis d'après l'observation de résultats préliminaires. Ces intervalles sont établis de façon empirique, certains étant imposés par les limites de la machine. Dans le chapitre 5, nous verrons l'influence de ces paramètres sur les résultats et sur la recherche de notre solution; nous présenterons également une analyse de sensibilité.

Tableau I

Paramètres de simulation

Paramètre de la simulation	Notation	Intervalle
Nombre d'individus par population	N_p	[100 ; 2000]
Nombre de générations	N_g	[100 ; 3000]
Pourcentage d'élitisme	p_e	[2 ; 20]
Probabilité de crossover	p_c	[0.9 ; 1]
Probabilité de mutation	p_m	[0 ; 0.2]
Profondeur maximale à la 1ère génération (niveaux)	max_prof	[2 ; 4]
Longueur maximale (nœuds)	N_{n_max}	[15 ; 30]

Il faut noter qu'à ces paramètres vient s'ajouter le type de modèle de régression utilisé pour mesurer la corrélation entre les données de terrain et les données hyperspectrales.

4.6 Conclusion

Il existe de nombreuses approches possibles en programmation génétique et les choix à faire dépendent en grande partie de l'application. Nous venons de décrire les principaux éléments qui interviennent lors de nos simulations et il nous reste à présenter les résultats, à les interpréter, à les discuter, pour finir par une série de recommandations permettant d'approfondir davantage cette étude.

CHAPITRE 5

RÉSULTATS : DISCUSSION ET INTERPRÉTATION

5.1 L'analyse d'images hyperspectrales

Lorsque les images hyperspectrales ont été acquises, corrigées et calibrées, elles doivent être analysées. Les trois méthodes les plus classiques pour l'étude du contenu de ces images sont a) la sélection de bandes; b) les indices de végétation et c) le « spectral unmixing » qui pourrait se traduire par le démêlage spectral.

En imagerie hyperspectrale, il est fréquent d'être en possession de plusieurs centaines de bandes spectrales dont de nombreuses sont hautement corrélées et contiennent une information redondante. Il existe alors de nombreuses techniques comme l'analyse par régression multiple, l'agrégation (« clustering ») ou encore l'analyse discriminante pour réduire cette quantité de bandes en un sous-ensemble contenant l'information essentielle.

Les indices de végétation (IV), comme nous l'avons déjà noté à plusieurs reprises, sont de simples combinaisons de bandes spectrales, faciles à implémenter, qui permettent de mettre en exergue certaines caractéristiques intrinsèques de la scène observée.

Les pixels d'une scène sont fréquemment constitués de plusieurs matériaux inconnus, dans des proportions inconnues (cf. figure 3 et l'article de Ross et al. [56]). Le démêlage spectral associé à la création de librairies spectrales fournit un outil puissant pour évaluer ces deux inconnues. En général, il faut commencer par trouver des pixels pures appelés « endmembers » (ne contenant qu'un seul matériau). Il est alors possible de créer une librairie spectrale au moyen de ces points et par la suite d'utiliser un algorithme pour trouver la composition des pixels de l'image.

Notre approche permet, par la recherche « libre » d'IV avec comme seules contraintes une longueur de solution maximale et le choix d'opérateurs arithmétiques simples, d'associer en quelque sorte les deux premières techniques précitées; en effet, notre algorithme va à la fois sélectionner les bandes discriminantes et constituer un IV adéquate avec celles-ci. C'est une différence fondamentale avec l'approche classique de détermination des IV.

5.2 Stratégie d'expérimentations

5.2.1 Base d'entraînement et base de validation

Comme décrit à la section 3.2 par la relation (3-1), les valeurs utilisées comme référence au sol (« ground truth ») utilisent deux types de mesures qui sont le ISF et les mesures SPAD; de ce fait, nous sommes limités à 88 données pour nos ensembles d'entraînement B_e et de validation B_v. Comme cela représente peu de valeurs, nous décidons de scinder aléatoirement ces données dans les proportions ¾ vs. ¼, ce qui donne les tailles définies à la relation (5-1) :

$$\text{taille}\left[B_e\right] = 66 \text{ et } \text{taille}\left[B_v\right] = 22 \tag{5-1}$$

Ainsi, nous allons entraîner notre algorithme sur 66 données et nous validerons les solutions trouvées sur une base de 22 données. À ce stade, il est important de noter que les tailles des bases utilisées sont petites comparativement aux applications classiques d'algorithmes évolutionnaires; dans notre cas, le but est de trouver un modèle performant pour le champ considéré et en aucun cas un modèle universel, même si nous

n'excluons pas la validité de nos solutions sur d'autres données. L'analyse des résultats permettra de conclure sur l'adéquation de l'usage de bases de petite taille.

Concernant la terminologie, nous utilisons le terme « validation » pour qualifier la base de données non apprise sur laquelle nous évaluons les solutions trouvées; elle peut être vue également comme une base de test puisque nous ne prenons pas de décision par rapport à l'erreur observée sur cette base, nous ne faisons que l'observer en arrêtant les simulations à la $3000^{ème}$ itération.

5.2.2 Données de référence

Dans la relation (3-1), comme nous ne connaissons pas la valeur constante $\rho s_{feuille}$, et que la valeur de la résolution de l'image peut changer selon le type d'image hyperspectrale utilisée, la valeur prise comme référence représentative de la quantité d'azote au sol, lors de l'entraînement et de la validation, est donnée par la relation (5-2).

$$V_{ref} = y = \frac{QN}{\rho s_{feuille} \times Aire_{pixel}} = \frac{QN}{constante} \qquad (5\text{-}2)$$

Cette valeur V_{ref} représente donc la quantité totale d'azote contenue dans la zone du champ couverte par le pixel considéré, au facteur ($\rho s_{feuille} \times Aire_{pixel}$) constant près. Cela est donc parfaitement équivalent à considérer les quantités d'azote puisqu'il y a un simple lien de proportionnalité entre ces deux types de valeurs.

Les régressions se font donc entre V_{ref} et les valeurs de la combinaison de bandes, sur la base d'entraînement et par la suite, les performances du modèle de régression trouvé sont évaluées sur la base de validation.

5.3 Choix des paramètres de simulation

Comme nous l'avons vu dans le tableau I, nous avons sept paramètres distincts pour chaque simulation avec pour chacun une gamme de valeurs étendue. Une analyse de sensibilité préliminaire permet de faire plusieurs remarques quant à la sélection de nos paramètres.

Nous avons choisi de mettre une pression relativement importante sur la taille des solutions pour les raisons évoquées à la section 4.4. Nous gardons des valeurs classiques pour les paramètres d'élitisme p_e, de crossover p_c et de mutation p_m dans le but de ne pas contraindre davantage le processus d'évolution. Ainsi, nous utilisons un pourcentage élevé de crossover aux alentours de 98% et un faible pourcentage de mutation aux alentours de 2%. Pour l'élitisme, une proportion de 2% est utilisée, encore une fois dans le but de donner un maximum de chance et de « place » au processus génétique.

Concernant la taille de la population et le nombre d'itérations, ici encore, nous choisissons de favoriser les croisements entre les individus et par conséquent, nous privilégions des simulations avec peu d'individus et beaucoup d'itérations. En effet, nous devons faire un compromis entre ces deux paramètres qui influencent directement le temps de calculs. L'idéal serait de considérer de grandes populations et de les faire évoluer à travers un nombre important d'itérations mais comme nous voulons faire plusieurs simulations avec un même jeu de paramètres, nous devons restreindre l'un de ces deux paramètres.

5.4 Caractéristiques des simulations et analyse de sensibilité

5.4.1 Comportement global

D'une façon générale, la valeur de fitness de la meilleure solution de chaque génération augmente de façon logarithmique, c'est à dire très rapidement au départ et lentement par la suite, tel qu'illustré par les graphiques 1, 2 et 3 de l'analyse de sensibilité. Ce comportement est caractéristique de la recherche d'un algorithme de PG.

5.4.2 Influence du modèle de régression utilisé

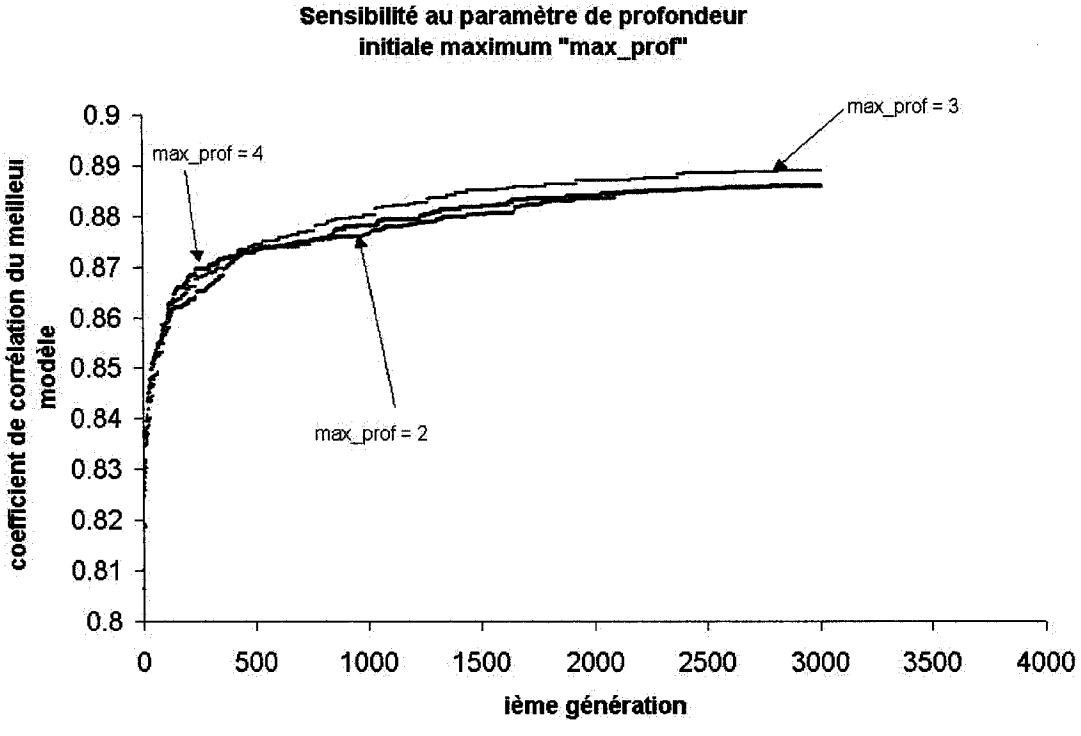
Lors des différents tests effectués avec les différents modèles de régression linéaire, exponentiel, puissance et logarithmique, les résultats obtenus ne diffèrent pas significativement. Toutefois, la régression logarithmique apporte dans la majorité des simulations des solutions légèrement plus performantes que les autres régressions et c'est donc avec ce type de modèle que nous allons effectuer l'analyse de sensibilité.

5.4.3 Choix du nombre d'itérations

Le choix du nombre d'itérations est partiellement relié à celui du nombre d'individus de la population. Il apparaît toutefois que le système se stabilise dans quasiment toutes les configurations avant d'arriver à la $3000^{ème}$ génération, après quoi, l'amélioration de la valeur de fitness du meilleur individu est relative au phénomène de surapprentissage. Ainsi, toutes les simulations seront conduites sur 3000 générations.

5.4.4 Influence de la profondeur à la première génération

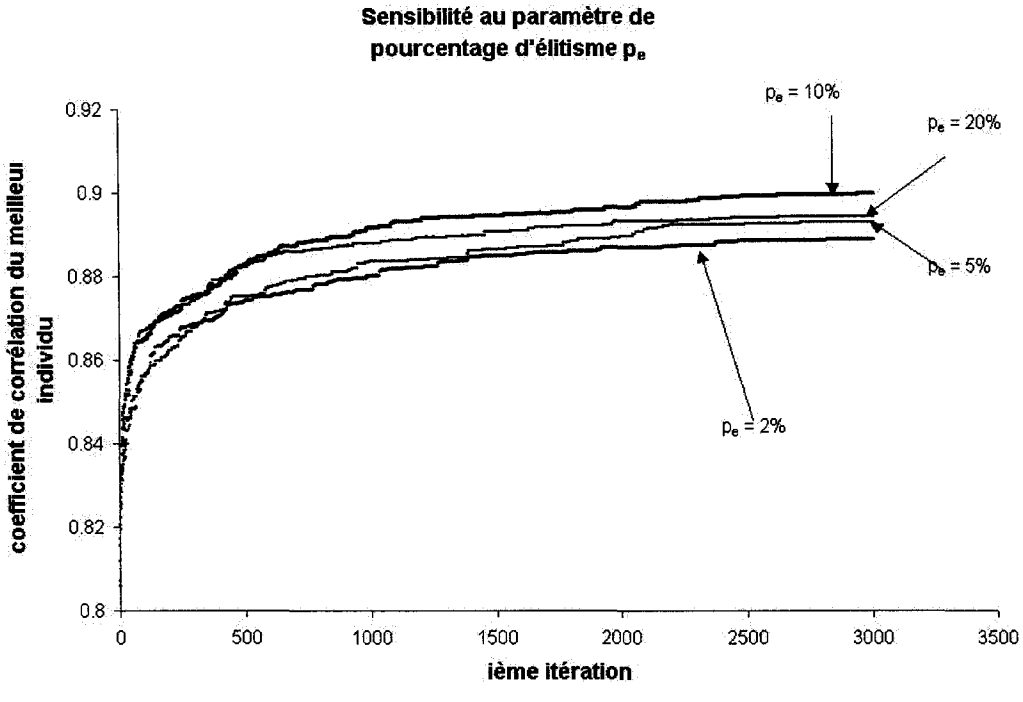
Le paramètre de profondeur maximale *max_prof* lors de la création de la première génération a pour but d'éviter l'explosion de la mémoire lors des premières itérations et de maximiser la diversité au sein de la génération initiale. Pour mesurer l'influence de ce paramètre lors de la recherche de notre solution, nous le faisons varier de 2 à 4 en gardant tous les autres paramètres fixes. Nous effectuons 7 simulations pour chacune des 3 valeurs de *max_prof* et nous obtenons les trois courbes illustrées au graphique 1. Nous voyons que les trois courbes sont très proches et que les meilleurs résultats sont obtenus lorsque ce paramètre est égal à 3. Ainsi, c'est cette valeur du paramètre qui sera utilisée.



Graphique 1 Sensibilité au paramètre max_prof

5.4.5 Influence de l'élitisme

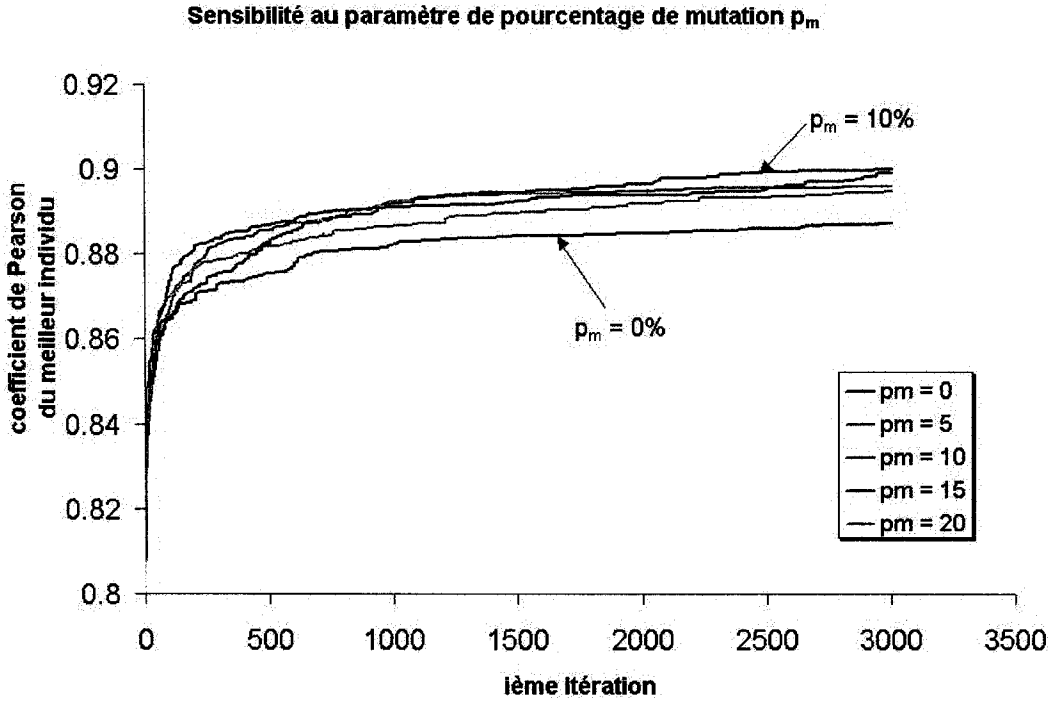
Le paramètre p_e définit la proportion d'individus les plus performants directement recopiés à la génération suivante, sans passer par les processus de crossover et de mutation. Sur le graphique 2, nous avons fait sept simulations pour des valeurs de p_e de 2%, 5%, 10% et 20%. Les courbes indiquent de meilleures performances pour p_e = 10%. Toutefois, comme pour le paramètre *max_prof*, nous voyons que ce paramètre n'est pas prépondérant.



Graphique 2 Sensibilité au paramètre d'élitisme p_e

5.4.6 Influence de la mutation

Le paramètre p_m définit la probabilité d'occurrence de la fonction de mutation sur les « enfants », en sortie du crossover. Ce paramètre est assez influent dans notre application puisqu'il permet d'apporter de la diversité au cours de la simulation. Nous constatons sur le graphique 3 que des 5 valeurs testées {0, 5, 10, 15, 20} (toujours sur 7 simulations), les meilleurs résultats sont obtenus pour p_m = 10%. Ici encore, nous constatons que ce paramètre n'est pas excessivement influent quant à la valeur du fitness du meilleur individu.

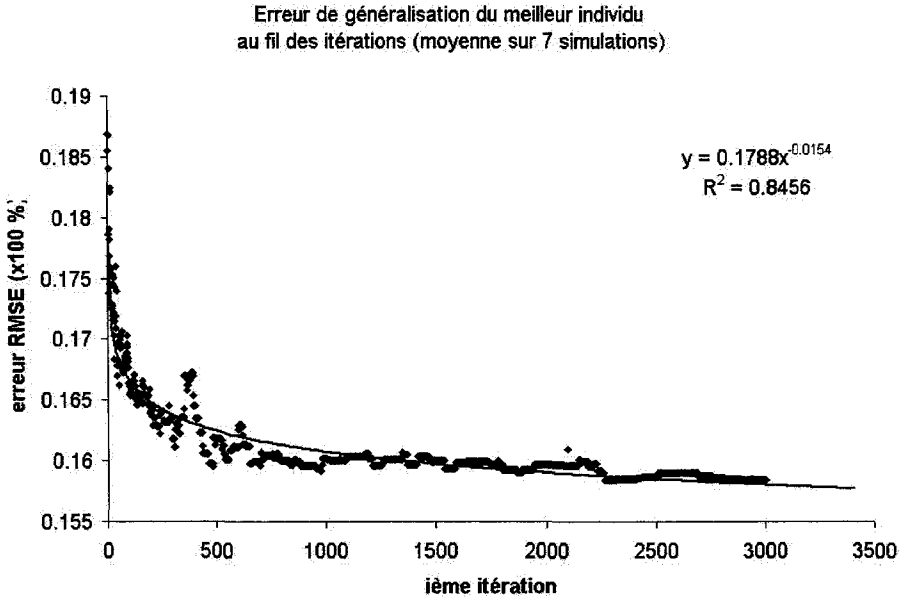


Graphique 3 Sensibilité au paramètre de mutation p_m

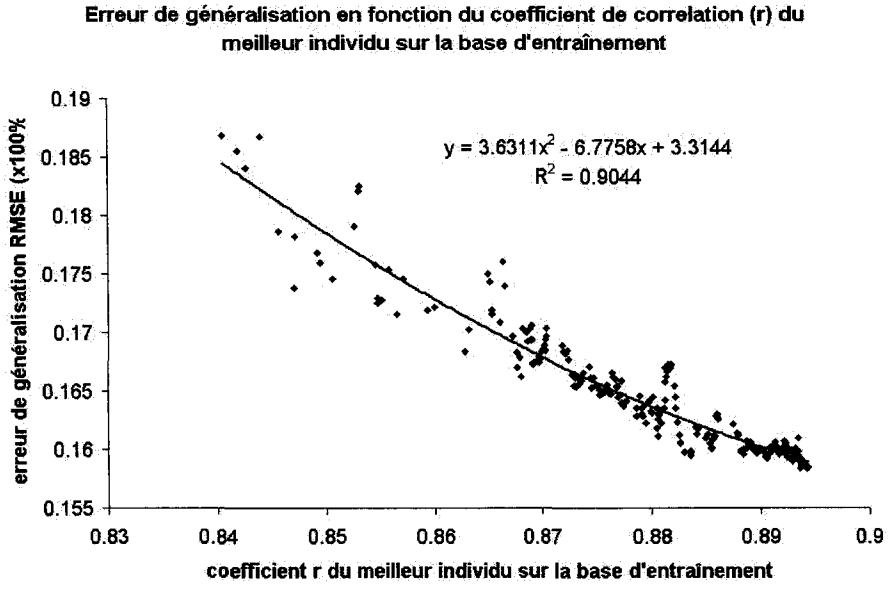
5.4.7 Erreur de généralisation

Compte tenu des précautions prises pour éviter le surapprentissage (explicitées à la section 4.4.2), nous introduisons une mesure d'erreur sur une base de test, disjointe de la base d'apprentissage; cette mesure représente l'erreur de généralisation et elle peut être vue comme un indicateur du surapprentissage au fil des itérations. Pour un processus classique, cette erreur de généralisation diminue jusqu'à un certain seuil au-delà duquel, elle commence à croître. C'est lorsque ce minimum est atteint que l'on doit arrêter la simulation puisque le modèle alors obtenu possède le meilleur pouvoir de généralisation.

Lors de 7 simulations, nous observons que l'erreur RMSE relative du meilleur individu ne cesse de diminuer sans recommencer à croître, tel qu'illustré au graphique 4. Cela dénote que notre stratégie pour éviter le surapprentissage est fructueuse.



Graphique 4 Erreur de généralisation du meilleur individu



Graphique 5 Erreur de généralisation en fonction du coefficient de corrélation du meilleur modèle sur la base d'entraînement

L'observation du graphique 5 permet de visualiser le lien entre l'erreur de généralisation du meilleur modèle de chaque itération et son coefficient de corrélation obtenu sur la base d'entraînement. Les deux bases étant totalement indépendantes, il est possible d'en conclure que nous évitons bien le phénomène de surapprentissage; en effet, la force de la corrélation du meilleur modèle sur la base d'entraînement ne mène en aucun cas à l'augmentation de l'erreur de généralisation sur la base de test, et ce jusqu'à la stabilisation du processus.

5.5 Mesures de performance

Pour évaluer les performances des IV trouvés, nous devons mesurer les caractéristiques des modèles de régression obtenus entre les indices et les données de terrain. Pour ce faire, en plus de l'évaluation des coefficients de détermination des modèles, nous utilisons plusieurs mesures d'erreurs répertoriées dans le tableau II. Nous mesurons ces

erreurs à la fois sur la base d'entraînement (erreur d'apprentissage) et sur la base de validation (erreur de généralisation). En plus, pour visualiser de façon claire la capacité de prédiction de l'IV et du modèle trouvés, nous comparons la valeur d'azote estimée V_e par le modèle à la valeur de référence V_{ref} en cherchant la pente α et le coefficient R^2 de la droite d'équation $y = \alpha * x$; une solution idéale doit avoir $\alpha \approx 1$ et le coefficient $R^2 \approx 1$.

Tableau II

Mesures d'erreur utilisées pour l'évaluation des performances

$RMSE_{abs}$	$RMSE_{\%}$	NMSE
$RMSE_{abs} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$	$RMSE_{\%} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2}}$	$NMSE = \frac{\sum_{i=1}^{i=n}(\hat{y}_i - y_i)^2}{n \times \sigma^2_{data}}$

Avec :

- $\hat{y}$: valeur d'azote estimée par le modèle.
- y : valeur d'azote réelle mesurée.
- σ^2_{data} : variance des mesures d'azote.
- n : taille de l'ensemble considéré (entraînement ou validation).

La racine carrée de l'erreur quadratique moyenne absolue ($RMSE_{abs}$) donne une estimation de l'erreur du modèle dans l'unité de mesure d'azote tandis que la $RMSE_{\%}$ est normalisée par rapport aux mesures réelles et donne donc une erreur relative en pourcentage, plus simple à interpréter. Concernant l'erreur quadratique moyenne normalisée (EQMN ou NMSE), la moyenne des carrés des erreurs est normalisée par la variance totale des erreurs et par conséquent, seule une valeur inférieure à 1 indique une

prédiction meilleure que la simple utilisation de la moyenne; plus cette valeur est basse, meilleure est la prédiction de ce modèle.

5.6 Performances des IV de la littérature pour l'évaluation de l'azote

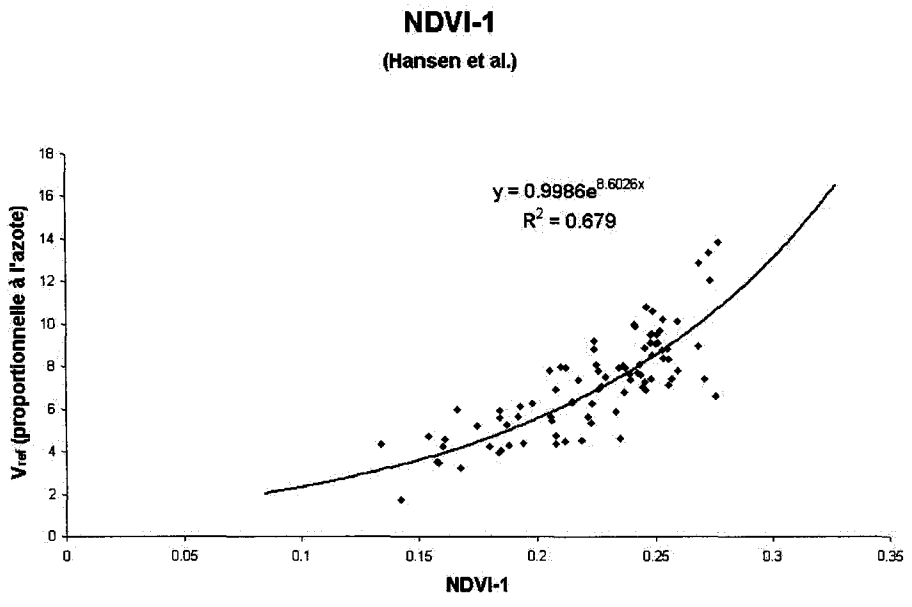
5.6.1 Les NDVI de Hansen et al.

Pour mesurer la performance de nos résultats, nous les comparons aux indices proposés dans la littérature. Hansen et al.[32] dans leur étude ont cherché un indice ayant la forme d'un indice de végétation normalisé (cf. la relation (2-1) à la section 2.4.2) qui représente la variabilité en azote au sein d'un champ de blé. Comme nous l'avons décrit dans la section 2.4.2 précitée, les meilleurs résultats obtenus ont indiqué une corrélation exponentielle avec un coefficient $R^2 = 69\%$.

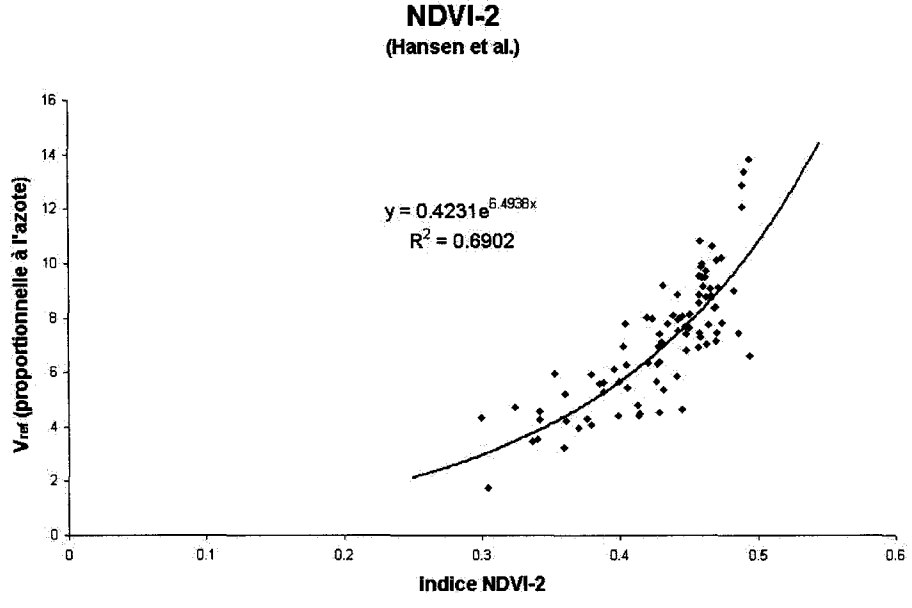
Sur nos données les trois indices proposés par Hansen et al. donnent des résultas presque parfaitement identiques tel qu'illustré aux graphiques 6, 7 et 8 ci-après.

Hansen et al. rapportent des coefficients R^2 de 69% pour des modèles de régression exponentielle pour ces trois indices NDVI-1, NDVI-2 et NDVI-3 et nous retrouvons des valeurs oscillant entre 67.90% et 69.02%. Nous allons donc prendre ces résultats comme référence pour évaluer la performance des nôtres.

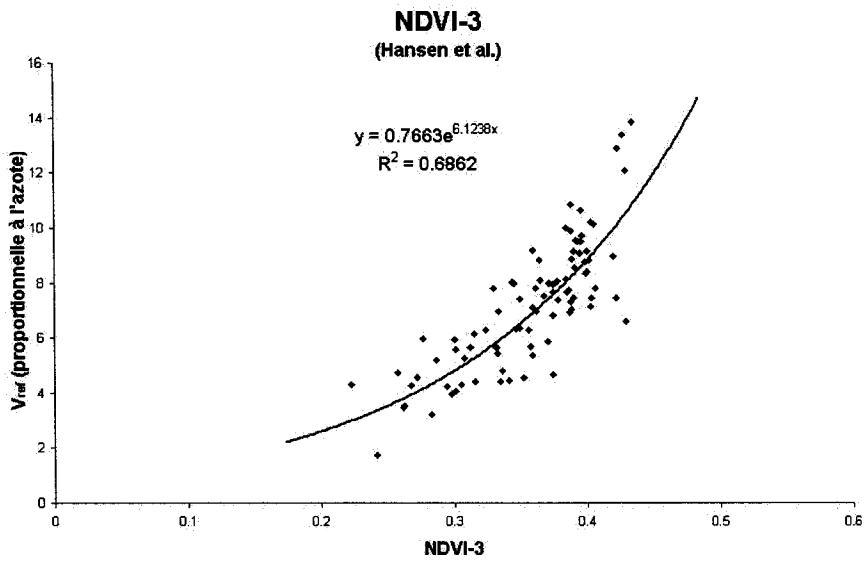
Nous allons également reprendre les indices de végétation classiques, parfois utilisés pour évaluer l'azote dans un champ.



Graphique 6 Indice NDVI-1



Graphique 7 Indice NDVI-2



Graphique 8 Indice NDVI-3

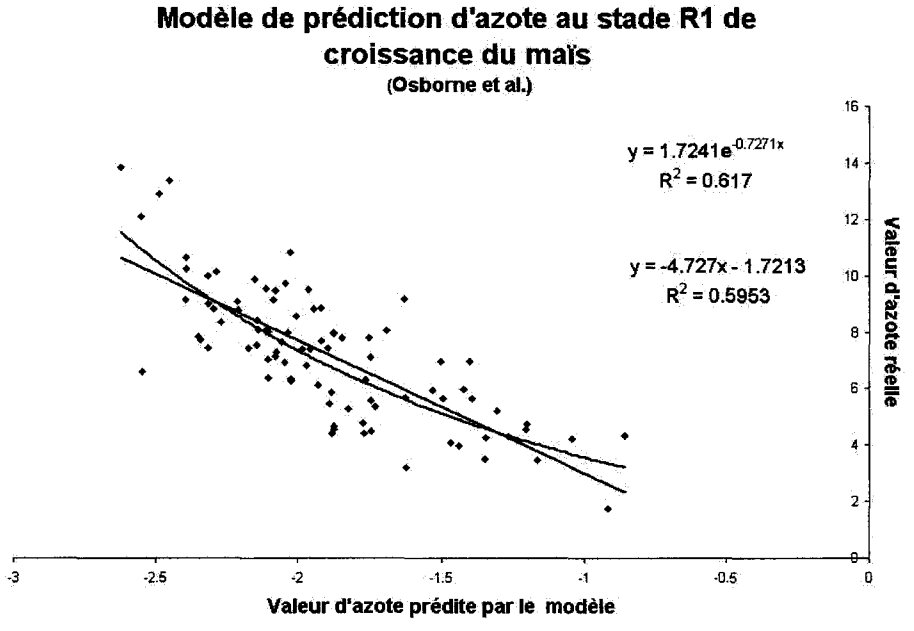
5.6.2 Modèle de Osborne et al

Dans leur étude, Osborne et al. [70] ont effectué des régressions multiples entre des données spectrales de réflectance aux bandes i (R_i) et des mesures d'azote et ont trouvé pour le stade R1 de la croissance du maïs (période d'acquisition de nos données cf. section 3.1), la valeur QN d'azote décrite par la relation (5-3).

$$QN = 0.884 + 0.760 \times R_{430} - 0.973 \times R_{435} + 0.277 \times R_{535} - 0.099 \times R_{760} + 0.054 \times R_{965} \quad (5\text{-}3)$$

Sur leurs données, ce modèle de prédiction du contenu de l'azote possède un coefficient de détermination $R^2 = 82\%$.

Ne possédant pas une gamme spectrale aussi large (cf. ANNEXE 4), nous négligeons le dernier terme et nous obtenons le graphique 9. Ainsi, sur notre ensemble de données, la sortie du modèle d'Osborne et al. est corrélée de façon exponentielle aux mesures d'azote avec un coefficient $R^2 \approx 62\%$.



Graphique 9 Modèle de prédiction de l'azote (Osborne et al., 2002)

Si au lieu de négliger le dernier terme R_{965} nous le remplaçons par la valeur de réflectance du proche infrarouge ayant la longueur d'onde la plus élevée de nos données R_{928}, les résultats obtenus sont moins bons que ceux présentés au graphique 9.

5.6.3 Autres indices classiques

Dans cette partie, nous ne ferons que mentionner les performances des autres indices pour ne pas alourdir le contenu. Aucun des autres IV testés n'améliore les résultats des deux travaux de Hansen et al. et de Osborne et al. sur notre ensemble de données, tel qu'illustré par le tableau III, dans lequel aucun des coefficients R^2 n'est supérieur aux résultats précités.

Tableau III

Coefficient R^2 des IV classiques, évalués sur nos données d'étude

NDVI: $\frac{R_{800} - R_{670}}{R_{800} + R_{670}}$	**RDVI**	**SR**	**MSR**	**SAVI**	**OSAVI**	**SARVI**
0.5635	0.5459	0.5650	0.5651	0.5630	0.5633	**0.5890**

5.6.4 Bilan des indices et modèles de la littérature

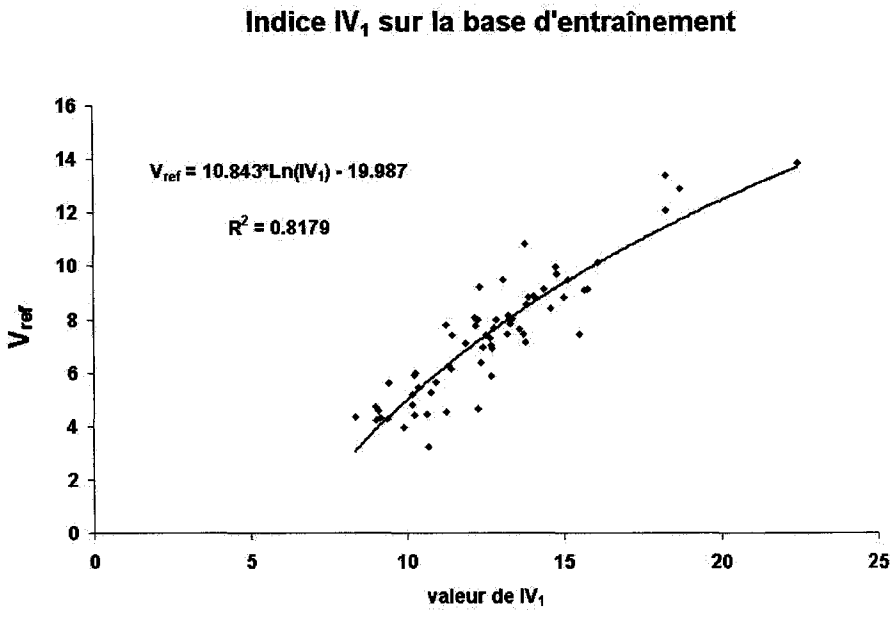
Sur nos données d'études, le modèle de la littérature qui donne les meilleurs résultats est la NDVI-2 de Hansen et al [32]. Cet indice possède les caractéristiques présentées au graphique 7 et résumées dans le tableau IV. Nous allons maintenant présenter la meilleure solution trouvée par notre algorithme.

5.7 Présentation du meilleur résultat trouvé par notre algorithme

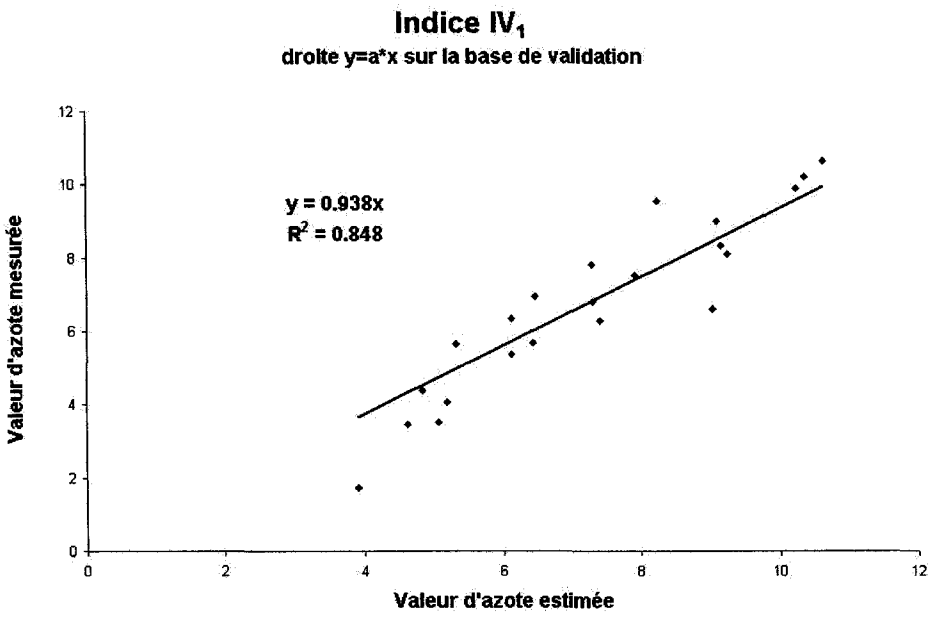
La solution IV_1 qui a montré les meilleures performances, lors des simulations effectuées, possède la syntaxe donnée par la relation (5-4), dans laquelle b_i désigne la bande spectrale i du capteur CASI, conformément à ce qui est décrit au tableau VII placé en annexe.

$$IV_1 = \frac{b_3 \times b_8 \times (b_{10})^3 \times b_{56}}{b_{31} \times (b_{37})^2 \times b_{70} \times (b_{34} - b_{63} + b_{47})} \tag{5-4}$$

Cet indice de végétation est corrélé de façon logarithmique aux valeurs d'azote V_{ref} mesurées en champ avec un coefficient $R^2 = 81{,}79\%$, comme montré au graphique 10.



Graphique 10 Régression logarithmique entre IV_1 et V_{ref}



Graphique 11 Performance du modèle sur la base de validation

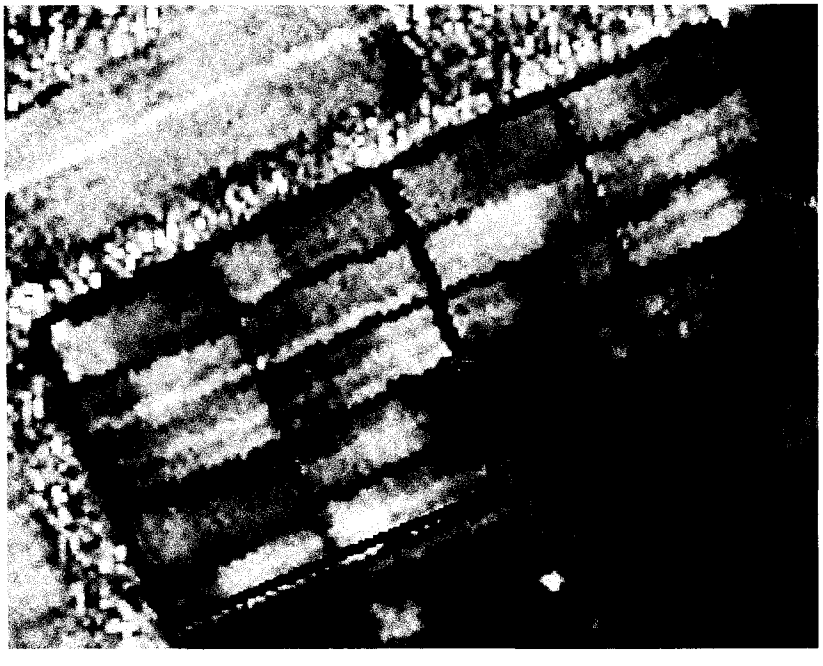


Figure 16 Image de IV_1 en tons de gris

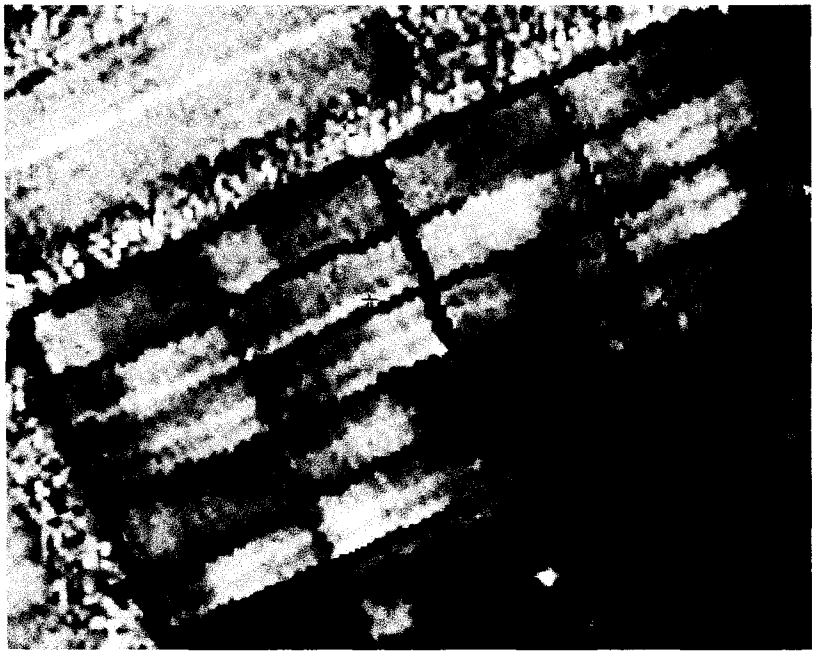


Figure 17 Carte de l'azote du champ à partir du modèle associé à IV_1

Ainsi, pour avoir une estimation de l'azote contenue dans le champ, il faut considérer les valeurs des tons de gris de l'indice IV_1 de la figure 16 comme variable du modèle logarithmique $V_{ref} = f_{reg}(IV_1)$ figurant dans le graphique 10. Nous obtenons alors une carte d'estimation de l'azote du champ (figure 17), avec les performances répertoriées dans le tableau V.

L'observation de la figure 17 permet de visualiser la variabilité en azote au sein du champ; cette figure nous montre qu'au niveau des parcelles ayant subi les mêmes traitements en azote et en herbicides, il existe une variabilité importante en azote; cela met en évidence l'hétérogénéité du champ en matière d'absorption d'azote par les plants. Toutefois, en comparant la figure 17 au plan d'expérience décrit à la figure 10, au sein de chaque groupe de trois parcelles adjacentes (ayant subi le même traitement en herbicide, Wi) nous sommes en mesure d'identifier lequel des trois traitements (N_0, N_1 et N_2) chacune d'elle a subi; en effet, la parcelle ayant subi le traitement N_2 apparaît

toujours comme celle globalement plus claire (donc plus d'azote) à la figure 17, tandis que celle ayant subi le traitement N_0 est toujours globalement la plus foncée.

5.8 Bilan et comparaison des résultats

L'indice IV_1 trouvé lors de notre étude est composé de bandes se trouvant dans les domaines du violet (1), du bleu (2), du rouge (3) et du moyen infrarouge (4). Au total, il est composé de dix bandes différentes et fournit une estimation de l'azote du champ plus précise que celle du meilleur indice trouvé dans la littérature. Les résultats comparatifs sont présentés dans le tableau IV pour la base d'entraînement et dans le tableau V pour les performances sur la base de validation; nous faisons un comparatif séparé sur ces deux sous-ensembles distincts de données pour que cette comparaison ait du sens.

Tableau IV

Performances de IV_1 et NDVI-2 sur la base d'entraînement

Solution	R^2	Nombre de bandes différentes	NMSE	$RMSE_{abs}$[10]	$RMSE_{\%}$[11] (x100)	Pente (α) de la droite $y = \alpha * x$	R^2 de la droite $y = \alpha * x$
IV_1	**0.818**	10	**0.179**	**0.981**	**0.127**	**0.984**	**0.781**
IV_1 sur W1			0.092	0.868	0.102		
IV_1 sur W2			0.268	1.112	0.148		
IV_1 sur W3			0.222	1.161	0.145		
IV_1 sur W4			0.216	0.704	0.105		
NDVI-2	0.647	**2**	0.345	1.372	0.177	1.033	0.655

[10] désigne la racine carrée de l'erreur quadratique moyenne absolue, exprimée dans l'unité des mesures.
[11] désigne la racine carrée de l'erreur quadratique moyenne relative en pourcentage.

Tableau V

Performances de IV_1 et de NDVI-2 sur la base de validation

Solution	NMSE	$RMSE_{abs}$	$RMSE_{\%}$ (x100)	Pente (α) de la droite $y = \alpha * x$	R^2 de la droite $y = \alpha * x$
IV_1	**0.183**	**1.020**	**0.143**	0.938	**0.848**
IV_1 sur W1	0.254	1.215	0.149		
IV_1 sur W2	0.170	0.874	0.131		
IV_1 sur W3	0.072	0.724	0.093		
IV_1 sur W4	0.309	1.303	0.235		
NDVI-2	0.289	1.283	0.188	**0.977**	0.702

L'analyse des tableaux 4 et 5 indique de meilleures performances de l'indice IV_1 sur NDVI-2, et ce avec presque toutes les mesures utilisées sur les deux bases. Le comparatif effectué sur la base d'entraînement a peu de sens puisque ces données ont été apprises par notre indice. Toutefois, sur la base de validation inconnue des deux indices, les résultats sont exploitables et permettent de conclure que notre approche fournit de meilleures performances que celles proposées dans la littérature.

En observant les erreurs obtenues sur les parcelles en fonction du traitement en herbicide Wi, nous observons que les données de W1 et de W4 ont été bien apprises par le modèle (avec respectivement des $RMSE_{\%}$ de 10.2 % et 10.5%) tandis que les données de W2 et W3 ont des erreurs d'apprentissage plus élevées. Sur la base de validation, c'est le phénomène opposé qui est observé; le modèle généralise mieux sur W2 et W3 que sur W1 et W4. Cela vient confirmer le fait qu'il existe une grande hétérogénéité artificielle entre les parcelles ayant subi des traitements en herbicides différents, ce qui complique certainement l'apprentissage comparativement au cas d'un champ réel.

En conclusion, nous avons mis en évidence la meilleure solution qui soit ressortie de nos simulations et nous avons vu que les bandes spectrales impliquées dans l'indice de végétation IV_1 ne contiennent pas de bande spectrale dans le domaine du vert; l'information pertinente pour quantifier l'azote a été trouvée dans les domaines du bleu, du violet, du rouge et de l'infrarouge, ce qui n'était pas nécessairement intuitif au seul examen des courbes de réflectance. Cela confirme que les variations spectrales impliquées par les différents niveaux d'azote ne peuvent pas être déduites directement des résultats observés à l'échelle foliaire lorsque nous considérons des données aériennes. L'utilisation de techniques d'apprentissage s'avère alors pertinente pour aller chercher des variations subtiles sur l'ensemble du spectre de réflectance.

Pour finir, le dernier chapitre est une étude du transfert technologique de notre méthode appliquée à la caractérisation de la variabilité en azote dans la canopée d'un champ de maïs. L'aspect financier n'est pas traité mais pourrait faire l'objet de travaux futurs.

CHAPITRE 6

TRANSFERT TECHNOLOGIQUE

Dans ce chapitre, nous proposons une analyse dont le but est de déterminer le nombre minimal de relevés en azote nécessaire à l'entraînement de notre algorithme, pour la détermination d'une solution pertinente. Concrètement, les relevés in-situ peuvent être effectués par l'agriculteur lors de son « tour de plaine » et par conséquent, il est utile de connaître le nombre minimal acceptable de ces mesures afin de réduire cette tâche. Les données de terrain sont coûteuses (service d'une compagnie spécialisée, tests en laboratoire, temps d'acquisition...) et cela justifie que nous nous intéressions à en diminuer la quantité nécessaire.

D'une façon générale, il est important que les données récoltées par l'agriculteur aient des valeurs étendues et représentent au mieux la variabilité globale du champ en azote. En effet, la qualité de généralisation du modèle trouvé par notre algorithme dépend en très grande partie de la qualité des données utilisées durant la phase d'apprentissage. Ceci n'est pas un problème pour un producteur agricole qui connaît généralement très bien les caractéristiques intrinsèques de ses parcelles et qui est donc capable de faire ses relevés de façon adéquate au regard de cette exigence.

D'un point de vue pratique pour l'étude ci-après, nous composons volontairement notre ensemble d'apprentissage avec des données couvrant l'étendue de la plage de variations de l'azote du champ. La base de test qui sert au calcul de l'erreur de généralisation reste identique au cours de toutes les simulations pour permettre de comparer les résultats entre eux. Les paramètres des simulations utilisés sont les paramètres optimaux trouvés lors de l'analyse de sensibilité. Le modèle de régression utilisé est logarithmique.

Habituellement, les algorithmes qui utilisent des processus évolutionnaires nécessitent une quantité de données d'apprentissage importante pour produire de « bonnes solutions ». Le terme « bonnes solutions » dépend de l'application considérée et elles sont évaluées par l'erreur produite sur des données inconnues. Dans l'analyse ci-dessous, nos comparons les erreurs des meilleures solutions trouvées, pour des tailles d'ensembles d'apprentissage N={5, 10, 15, 20, 25, 30, 35, 40, 45}, avec l'erreur de généralisation produite par le modèle utilisant IV_1 (cf. chapitre précédent). Pour chacune des tailles N de l'ensemble d'apprentissage, nous observons également l'évolution de l'erreur de généralisation du meilleur individu au cours des itérations afin de prévenir le surapprentissage. Le meilleur individu est défini comme étant celui dont l'erreur de généralisation RMSE% est la plus faible; cinq simulations sont menées pour chaque taille d'ensemble d'apprentissage et la meilleure performance est rapportée dans le tableau VI ci-après.

Le premier constat est que l'erreur de généralisation augmente globalement lorsque la taille de l'ensemble d'apprentissage diminue; toutefois, cette augmentation n'est pas importante en comparaison du gain en nombre d'échantillons nécessaires à l'entraînement. De même, on remarque que le coefficient de corrélation de la meilleure solution, sur la base d'apprentissage, a tendance à diminuer lorsque le nombre d'échantillons d'apprentissage augmente; ceci se comprend facilement puisqu'il est plus simple de trouver un modèle qui apprenne correctement 5 valeurs que 40 valeurs.

Le nombre d'itérations nécessaire pour l'occurrence de la meilleure solution a tendance à augmenter avec la taille de l'ensemble d'apprentissage. Pour de petites tailles de base d'entraînement, le surapprentissage caractérisé par l'augmentation de l'erreur de généralisation survient rapidement, souvent même au cours de la première itération; ceci est explicable puisque les données sont peu nombreuses donc vite apprises. Il semble que le processus génétique ne prenne de sens qu'à partir de N = 40, puisqu'il faut alors une moyenne de 784 itérations pour parvenir à la meilleure solution de la simulation.

Tableau VI

Précision du modèle en fonction du nombre de données d'entraînement

Taille N de l'ensemble d'apprentissage	Erreur $RMSE_{\%}$ de généralisation de la meilleure solution	Coefficient de corrélation du meilleur individu sur la base d'apprentissage	Nombre d'itérations pour l'occurrence de la meilleure solution
N=5	0.162	**0.994**	67
N=10	0.149	0.960	1
N=15	0.137	0.895	1
N=20	0.155	0.935	20
N=25	0.131	0.942	1
N=30	0.133	0.917	1
N=35	0.140	0.884	50
N=40	0.139	0.931	784
N=45	**0.126**	0.908	913

Nous constatons que nous obtenons des erreurs de généralisation souvent inférieures à celle de l'indice IV_1 ($RMSE_{\%}$ = 14.3%) qui était entraîné sur un plus grand nombre de données (N=66); cela peut être interprété par le fait que certaines des données d'entraînement étaient mauvaises ce qui aurait pu biaiser l'apprentissage. Cela met donc en évidence le fait que peu de données sont nécessaires à un bon apprentissage, dans la mesure où ces données sont justes. De même que pour entraîner un classifieur il faut être certain de la validité des labels des données d'entraînement, notre approche par régression nécessite l'emploi de données de référence précises. Idéalement, les données recueillies par l'agriculteur devraient représenter des configurations diverses des variables biophysiques autres que celle étudiée, dans le but d'entraîner notre modèle à

isoler la variabilité spectrale recherchée des variations parasites. Encore une fois, la connaissance à priori de l'agriculteur est une information qu'il est nécessaire d'utiliser pour permettre l'acquisition de données de qualité. En définitive, les résultats obtenus indiquent que le nombre de données nécessaires à l'entraînement a moins d'influence sur le résultat que la qualité de celles-ci et qu'il est donc possible d'appliquer notre approche dans un contexte de production agricole.

CONCLUSION

La télédétection offre la possibilité d'acquérir rapidement et de façon non destructive, de grandes quantités de données, sur de vastes surfaces. Les progrès constants dans la conception de capteurs permettent d'améliorer continuellement la précision des outils et d'en démocratiser l'utilisation dans de nombreux domaines, notamment en agriculture de précision. Dans ce domaine d'application, le récent développement de l'imagerie hyperspectrale permet d'envisager de nouvelles approches de gestion des ressources, dans la mesure où l'on est capable d'extraire, de façon précise, l'information pertinente contenue dans ces données.

Plusieurs techniques ont été développées pour analyser le contenu des images hyperspectrales de télédétection, avec des avantages et des inconvénients selon l'application. Pour la détection de paramètres biophysiques au sein des cultures, les indices de végétation (IV) ont démontré qu'ils pouvaient s'avérer être d'excellents outils d'évaluation; dans le cas précis de la quantification de la variabilité en azote, la difficulté de discerner de faibles variations dans les réponses spectrales d'éléments d'une même espèce (le maïs dans notre étude) est accentuée par le bruit des données. Bien que les techniques de corrections des images (atmosphérique, radiométrique et géométrique) soient en progrès constant, il apparaît encore difficile de trouver des indices capables de décrire des variables biophysiques indépendamment de tout autre facteur variable (radiance, humidité...). En d'autres termes, il semble que l'on ne soit pas encore capable de se rendre suffisamment indépendant des conditions extérieures pour trouver un modèle généralisant la détermination d'une variable biophysique d'un champ, à partir de données hyperspectrales.

Pour palier à cette incapacité de généralisation, il est nécessaire de posséder une information supplémentaire sur la « réalité du terrain », au moment de l'acquisition des données de télédétection. En mesurant la variable biophysique d'intérêt en quelques

points du champ choisis de façon astucieuse, il est alors possible de procéder à une estimation précise de cette variable en tout point de la canopée.

Dans notre étude, nous avons conçu un algorithme empruntant les concepts de la programmation génétique pour élaborer un IV qui décrive le plus justement possible la variabilité en azote au sein d'un champ de maïs. Les résultats obtenus sur nos données ont montré la supériorité de la solution trouvée par notre approche, sur tous les indices et modèles de la littérature. Nous avons considéré le cas de l'azote au sein d'un champ de maïs et il est possible d'appliquer notre algorithme pour n'importe quelle autre variable biophysique (phosphore, ISF, teneur en eau...) et ce, sur n'importe quel autre type de culture (blé, soja...). En attendant que l'amélioration combinée des moyens d'acquisition et des méthodes de correction permette l'inférence de modèles « universels » pour un problème donné, sans recours à la réalité de terrain, ce type d'approches offre des performances exploitables en pratique.

RECOMMANDATIONS

Les premières recommandations concernent le type de données utilisées; l'étendue spectrale des données fournies par le CASI est [409 ; 932] nm alors que certaines études[12] rapportent des pics d'absorption de l'azote aux longueurs d'onde 1640 nm et 2100 nm dans le moyen infrarouge. Il serait par conséquent intéressant d'évaluer les performances de notre algorithme sur des données hyperspectrales couvrant cette partie du spectre électromagnétique, acquises par le capteur AVIRIS par exemple. Toutefois, ces capteurs coûtent plus chers car ils font appel à des technologies plus récentes pour ce qui est de la capture et de l'enregistrement de l'énergie électromagnétique dans le domaine du moyen infrarouge.

Concernant la quantité de données, il pourrait être judicieux de faire appel à des techniques de sélection ou d'extraction de caractéristiques afin de réduire la dimension de l'espace de recherche et de faciliter l'obtention d'une solution. L'Analyse en Composantes Principales (ACP) ayant été testée sans amélioration des résultats, d'autres méthodes comme la GLDB (« Generalized Local Discriminant Bases ») [71], la SPCT (« Segmented Principal components transformation ») [72] ou encore une technique empruntée à la poursuite de projection, proposée par Jimenez & Landgrebe [73] pourraient être testés pour réduire la dimension de l'espace des caractéristiques; toutes ces méthodes sont des adaptations d'outils mathématiques destinées à des applications utilisant des images hyperspectrales.

Au sujet des données de terrain, il serait préférable de travailler avec des mesures d'azote plus précises que les estimations que nous avons faites; de même, pour s'assurer une meilleure confiance dans les résultats obtenus, il serait nécessaire d'être en possession d'un plus grand nombre de ces mesures.

---

[12] http://webpages.acs.ttu.edu/smaas/asa2000/fitzgerald.htm

La suite des recommandations concerne l'amélioration de notre algorithme de recherche. Nous avons utilisé des fonctions binaires (les quatre opérateurs arithmétiques de base) pour construire les IV de nos populations; nous pourrions considérer l'ensemble des opérateurs unaires de transformation comme le logarithme naturel *ln(x)*, la fonction inverse *1/x*, la racine carrée $\sqrt{x}$ ou encore la fonction puissance x^{α}. En élargissant ainsi notre espace de recherche, il serait sans doute possible d'améliorer les performances de la solution trouvée.

Concernant la pression que nous avons exercée sur la longueur des individus par l'instauration d'un paramètre de longueur maximale, il serait intéressant de tester d'autres types d'opérateurs de crossover (un des responsables du « code bloat ») comme le « fair crossover » ou le « homologous crossover » [65, 67]; ces opérateurs permettent de réduire le « bloat » sans contraindre la recherche de façon trop radicale. Pour les autres opérateurs, nous pourrions faire de l'élitisme faible et tester d'autres méthodes visant à ne pas perdre de bons individus (garder systématiquement les deux meilleurs parmi les deux parents et les deux enfants). Nous pourrions également définir une fonction de coût qui pourrait être : $\text{Coût}(IV_1, IV_2) = \frac{\Delta R_{1,2}}{\Delta L_{1,2}}$; cette fonction permettrait d'évaluer le gain en précision $\Delta R_{1,2}$ d'un enfant IV_2 par rapport à son parent relativement à l'augmentation en longueur $\Delta L_{1,2}$, ceci menant à la décision de garder l'enfant ou bien un des parents. Cela nous obligerait alors à introduire un nouveau paramètre (valeur-seuil) pour déterminer la décision à prendre.

Concernant le maintien de la diversité au sein des populations, il serait pertinent d'implanter dans l'algorithme une mesure de diversité pour effectuer le suivi de cette caractéristique de façon précise.

Enfin, du point de vue de la validation, il serait utile, compte tenu du peu de données dont nous disposons, de faire de la validation croisée (« cross validation ») afin

d'évaluer plus rigoureusement les performances de notre méthode en exploitant au maximum nos données.

ANNEXE 1

Description générale de l'imagerie hyperspectrale

D'après les livres de Michel-Claude Girard [74] et de Thomas M. Lillesand [75]

Réflectance

Un objet éclairé par le soleil reçoit une énergie incidente directe E_I. L'objet qui reçoit cette énergie en absorbe une quantité E_A, en transmet une quantité E_T et en réfléchit la quantité E_R. Ce bilan est dépendant de la longueur λ et il est défini par l'équation (A-1) suivante :

$$E_I(\lambda) = E_R(\lambda) + E_A(\lambda) + E_T(\lambda) \qquad \text{(A-1)}$$

La réflectance d'un objet, d'un matériau est une caractéristique qui lui est propre et qui est indépendante des conditions externes. On la définit sous la forme d'un coefficient $\rho(\lambda)$, dépendant de la longueur d'onde, comme étant le rapport de l'énergie réfléchie à la longueur d'onde λ sur l'énergie incidente reçue comme montré par l'équation (A-2) suivante :

$$\rho(\lambda) = \frac{E_R(\lambda)}{E_I(\lambda)} \qquad \text{(A-2)}$$

La figure 2 de l'introduction montre l'allure de la courbe de réflectance pour l'ensemble des végétaux chlorophylliens. Seules de petites variations de cette courbe interviennent selon les caractéristiques spécifiques du végétal (composition chimique, teneur en eau...).

Bandes spectrales

Il est intéressant de posséder des mesures de réflectance pour des longueurs d'onde pertinentes, dépendant de ce que l'on observe. Dans le cas de la végétation, les domaines du spectre électromagnétique contenant l'information essentielle se situent dans le rouge et le bleu qui sont les zones d'absorption principale des pigments chlorophylliens, dans le vert pour la zone d'absorption des caroténoïdes et dans l'infrarouge pour évaluer la teneur en eau. Il est donc important de détecter la réponse de la scène sur l'ensemble du spectre électromagnétique allant du visible à l'infrarouge.

Les capteurs hyperspectraux (cf. la description ci-après) permettent de discrétiser la réflectance de la scène observée en bandes contiguës d'étendue L plus ou moins large en découpant ce spectre en portions B_i appelées bandes spectrales comme schématisé à la figure 18.

Chaque bande B_i représente la somme de l'énergie réfléchie par la scène, sur une portion du spectre de largeur L. Pour activer un détecteur (CCD) d'un capteur hyperspectral, il faut une énergie minimum. Plus L est élevée, plus l'énergie incidente (couvrant une grande zone du spectre électromagnétique) va être grande dans la bande considérée, offrant ainsi l'opportunité d'obtenir une taille de pixel plus petite (résolution plus fine). Il y a donc un lien de proportionnalité inverse entre la résolution spatiale et la résolution spectrale.

Ainsi, lors de l'utilisation d'un capteur hyperspectral, on obtient une image constituée par la superposition des réflectances de la même scène pour des longueurs d'onde différentes comme schématisé à la figure 19.

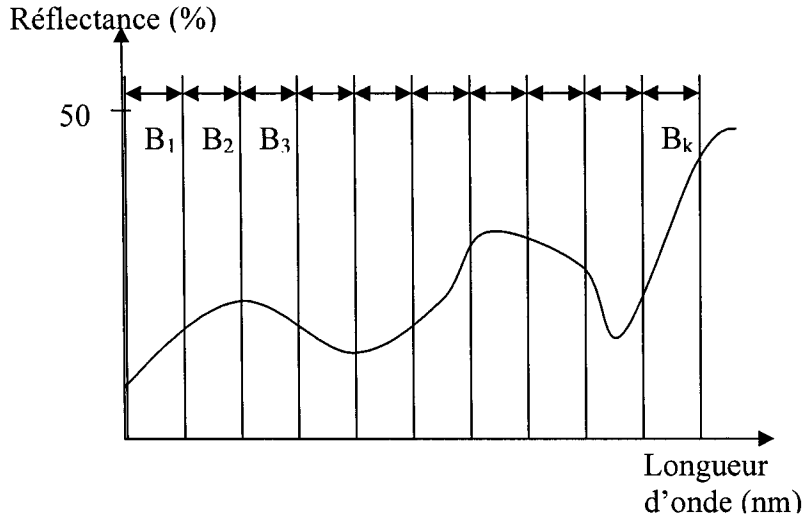


Figure 18 Représentation des bandes spectrales d'une scène quelconque

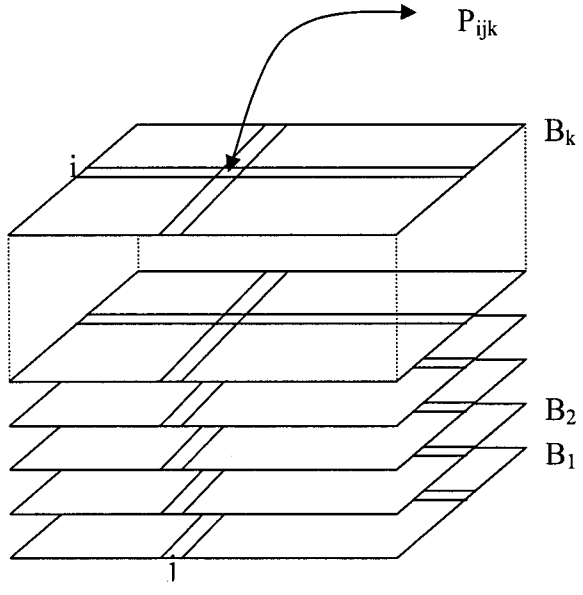


Figure 19 Structure d'une image hyperspectrale de k bandes

Chaque pixel P_{ij} de l'image hyperspectrale est donc défini par k coordonnées dans la base vectorielle des bandes spectrales $(B_1, B_2, ..., B_k)$ et peut donc être défini par le vecteur de caractéristiques montré en (A-3) :

$$P_{ij} = \begin{pmatrix} P_{ij1} \\ P_{ij2} \\ P_{ij3} \\ \cdots \\ P_{ijk} \end{pmatrix}_{(B_1, B_2, ..., B_k)} \qquad \text{(A-3)}$$

C'est ce formalisme qui est utilisé pour faire de la classification d'images hyperspectrales.

Les capteurs hyperspectraux

L'objectif de ce paragraphe est de présenter de façon générale le principe de fonctionnement des différentes catégories de capteurs multibandes embarqués à bord de satellites ou bien aéroportés. Les capteurs hyperspectraux permettent l'acquisition d'informations spectrales dans les domaines du visible (400nm à 700 nm) et de l'infrarouge proche, moyen réflectif et moyen (de 700 nm à 5 μm). Au-delà de ces longueurs d'onde, il y a le domaine de l'infrarouge thermique qui nécessite l'utilisation d'un radiomètre infrarouge. La principale différence de ces capteurs avec la technologie multispectrale est la largeur des bandes, beaucoup plus fine pour les capteurs hyperspectraux. Cela donne l'opportunité de détecter des variations plus sensibles dans le spectre de réflectance.

Il existe deux catégories de capteurs hyperspectraux a) le système à balayage; b) le système à barrettes. Tous ces capteurs répondent au même schéma de principe donné à la figure 20. Ces deux types de systèmes sont décrits brièvement dans les paragraphes suivants.

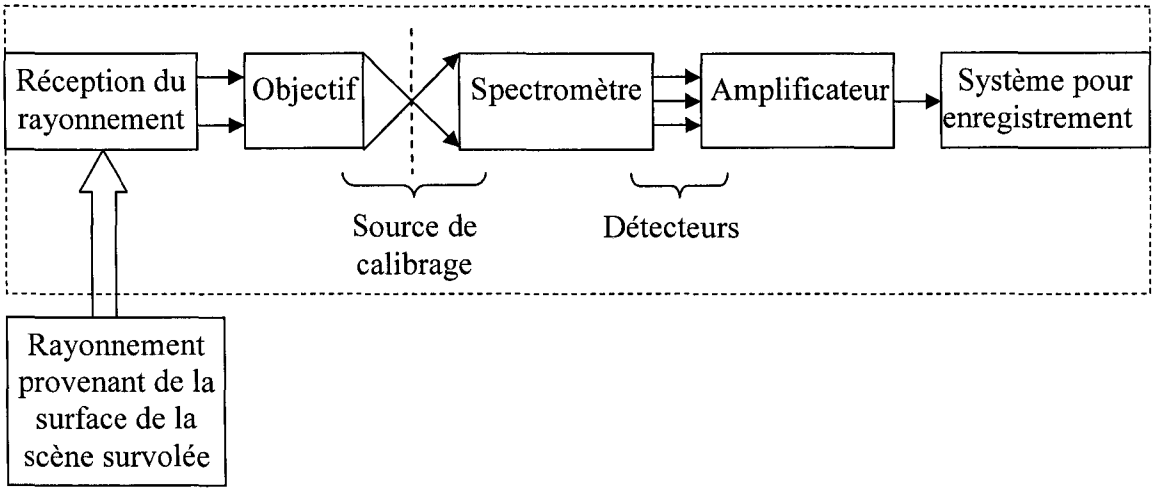


Figure 20 Schéma général d'un capteur numérique [74]

Système à balayage

Ce système qui est le plus ancien, équipait le satellite LANDSAT 1 en 1972. De nombreux capteurs aéroportés utilisent également cette technologie présentée à la figure 21.

C'est un miroir tournant ou oscillant qui va permettre de guider, vers un unique détecteur, l'énergie électromagnétique renvoyée par la surface investiguée. L'angle α de la figure 21 est appelé champ de vision instantané (en anglais, Instantaneous Field Of View, IFOV) et il est constant pour un capteur donné; l'altitude de vol et le IFOV

définissent la taille du pixel de l'image produite en sortie, c'est à dire la résolution géométrique du capteur.

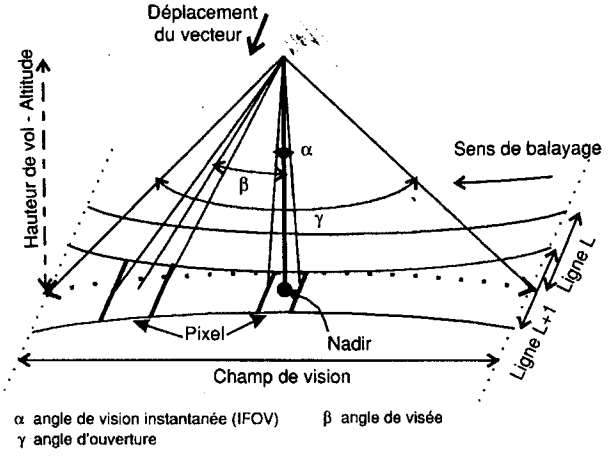


Figure 21 Schéma de principe d'un système à balayage

L'image est donc constituée de bandes dont la largeur est définie par α. L'adjacence de ces bandes et la couverture exhaustive de la scène sont assurées par la synchronisation parfaite de la vitesse et de l'altitude du vecteur (avion ou satellite) avec la vitesse de balayage du miroir. Lors de l'acquisition d'une image, la géométrie du capteur doit toujours être précisée. En effet, comme le montre la figure 21, il existe des déformations géométriques systématiques dépendant du IFOV lorsque la scène s'éloigne du Nadir. Ces déformations seront alors corrigées au moyen d'algorithmes spécifiques.

Système à barrettes

Ce système décrit à la figure 22, appelé aussi système en peigne, est plus récent que celui à balayage. Il est composé d'au maximum 12000 détecteurs CCD (Charge-Coupled Device) recevant simultanément l'information d'une bande de terrain survolée. À

chaque pixel correspond un détecteur CCD. Le gros avantage de ce système est qu'il permet un temps d'exposition t_e plus long; par conséquent, cela permet une diminution de la taille possible des capteurs du fait de l'augmentation de l'énergie incidente par unité de temps et donc une meilleure résolution spatiale. De plus, les performances radiométriques sont également meilleures avec cette technologie. Par exemple, pour le satellite SPOT, chaque milliseconde et demie, une bande de terrain de 10m sur 60km est couverte et discrétisée en 6000 pixels. C'est l'avancement du vecteur qui entraîne l'acquisition d'une nouvelle ligne et c'est pourquoi on appelle ce système également système « push broom » ou à ratissage.

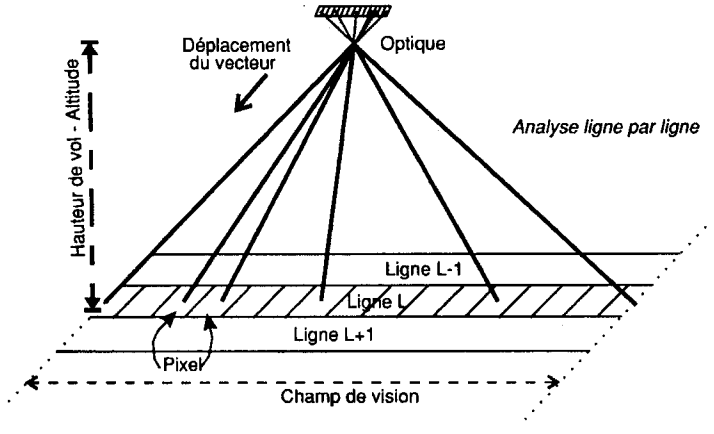


Figure 22 Schéma de principe d'un système à barrettes

Compact Airborne Spectrographic Imager (CASI)

Les données de l'étude proposée ont été acquises avec le spectromètre canadien CASI qui est de type à barrettes. Il est composé de 512 pixels spatiaux pour chacune des 288 bandes spectrales qu'il peut acquérir. La couverture spectrale s'étend du bleu 400nm au proche infrarouge 950nm. Les deux modes d'acquisition utilisés lors de la campagne GEOIDE 2000 sont :

a. 72 bandes spectrales de largeur 7.47nm avec une résolution spatiale de 2 mètres;
b. 7 bandes spectrales de 1.9nm avec une résolution spatiale de 1 mètre.

Dans le cas a, l'ensemble du spectre de réflectance est exploré tandis que dans le cas b, seule une sélection de bandes étroites est considérée; les 7 bandes ont été choisies pour leur pertinence dans le cadre de l'inspection d'un couvert végétal. Elles sont centrées sur les longueurs d'onde 489.5nm, 555nm, 624.6nm, 681.4nm, 706.1nm, 742.3nm et 776.7nm.

ANNEXE 2

Photo du champ-test et de la ferme MacDonald

Figure 23 Ferme MacDonald et champ-test en « L »

ANNEXE 3

Visualisation des données de l'étude

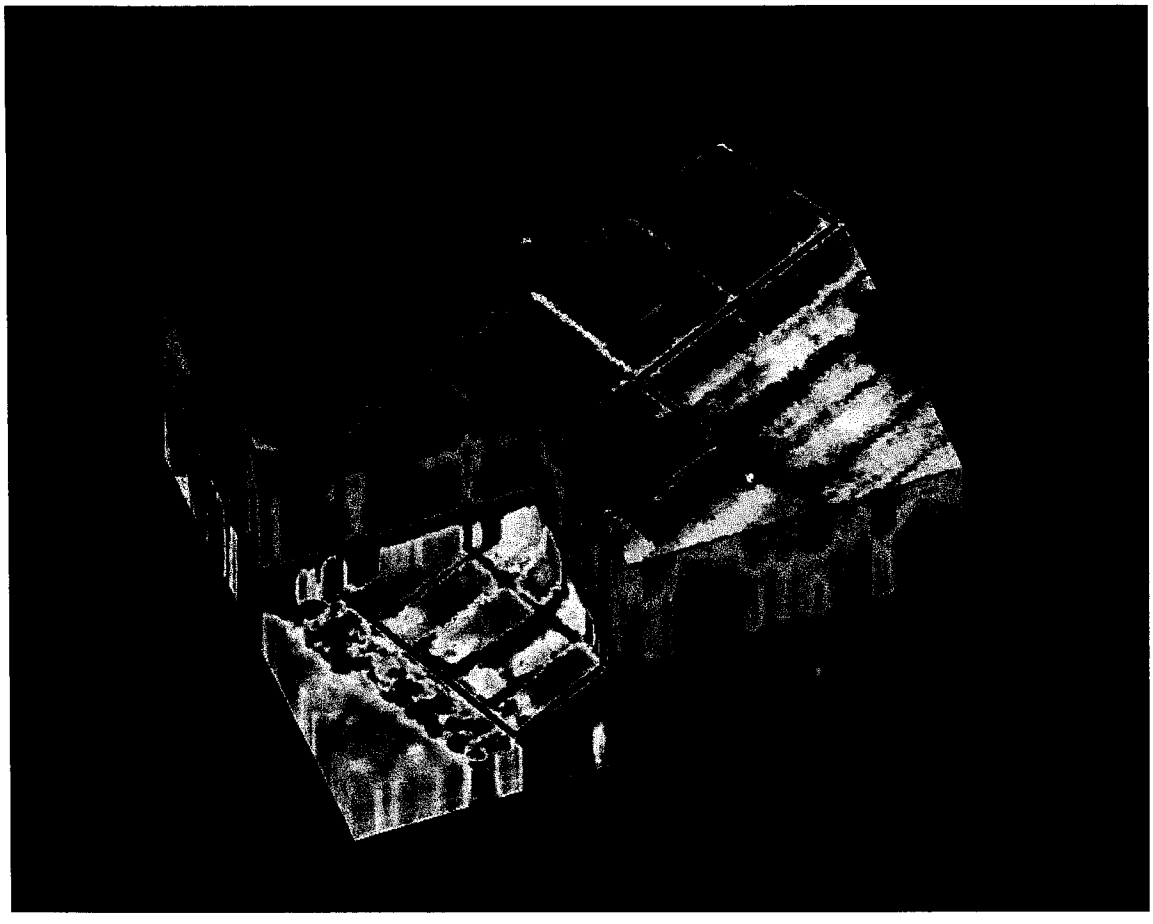


Figure 24 « Hypercube » des données spectrales, dans l'environnement de PCI.

Ce mode de visualisation superpose chaque bande spectrale de l'image hyperspectrale en appliquant une table de correspondance de pseudo-couleurs (PCT) aux valeurs de réflectance. Au sommet de l'hypercube, c'est la scène reconstituée (RGB) qui est visible afin de donner un repère visuel. Comme montré dans le coin inférieur gauche du cube, il est possible d'inciser le bloc pour, par exemple, repérer les bandes spectrales denses en informations.

On reconnaît ici en bleu le contour du champ en forme de « L » présenté à la figure 23.

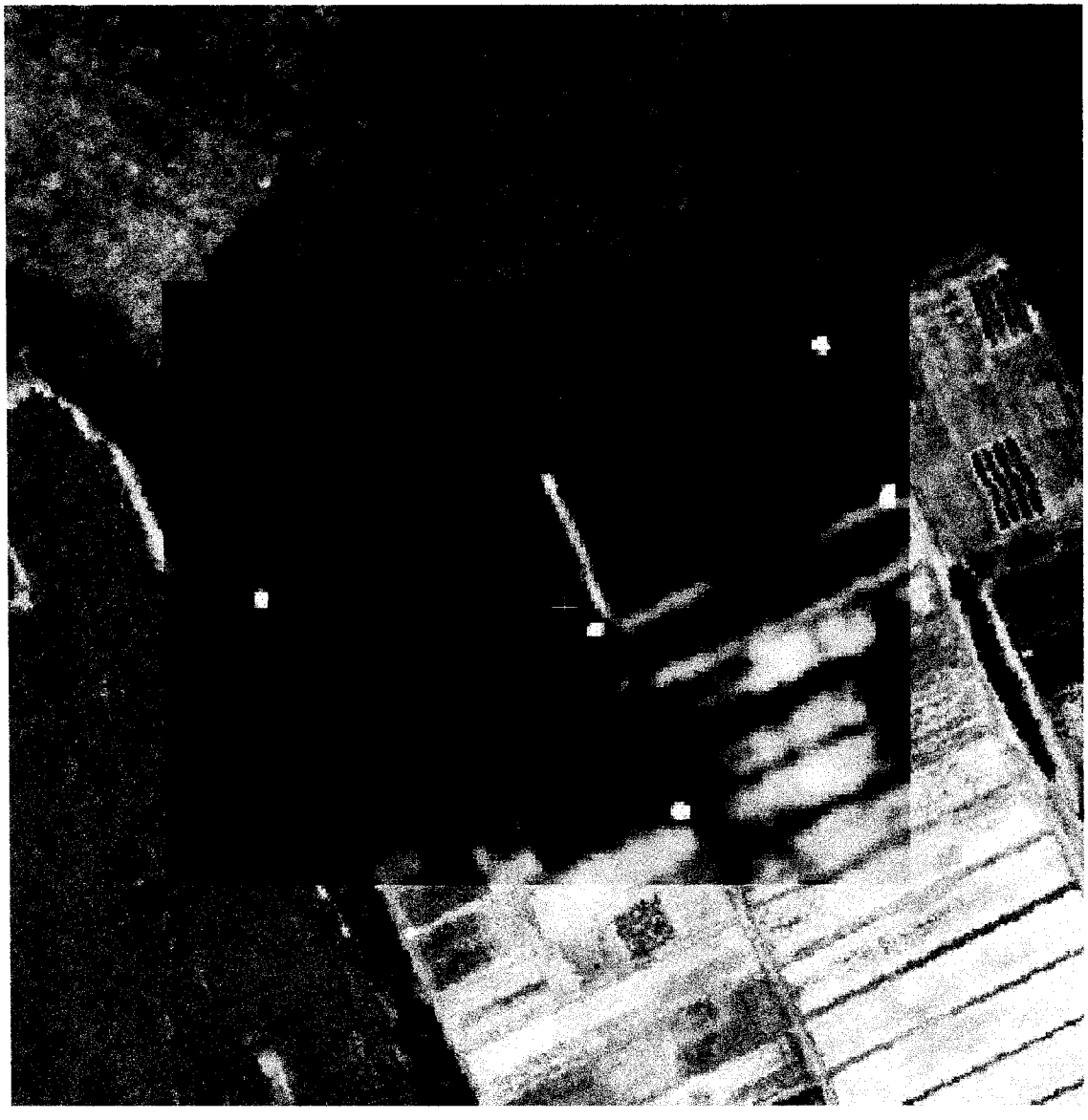


Figure 25 Image du champ-test, image réduite filtrée et points échantillonnés

Sur la figure 25, l'image recomposée (RGB) en haute résolution (1m x 1m) du champ-test est drapée par l'image moyennée basse résolution (2m x 2m) réduite, utilisée dans l'étude. Les marqueurs rouges sont les localités des mesures d'ISF effectuées en champ; c'est en ces points qu'a lieu l'échantillonnage des bandes spectrales.

ANNEXE 4

Caractéristiques spectrales du CASI

Tableau VII

Caractéristiques des bandes spectrales du CASI

CASI Sensor Range (nm)	409 - 947		
Number of Bands	72		
Wavelengths per band (nm)	7.472222222		
Band Number b_l	**Wavelength Range** (nm)		**Colour**
1	409	416.4722222	Violet
2	416.4722222	423.9444444	Violet
3	423.9444444	431.4166667	Violet
4	431.4166667	438.8888889	Violet - Blue
5	438.8888889	446.3611111	Blue
6	446.3611111	453.8333333	Blue
7	453.8333333	461.3055556	Blue
8	461.3055556	468.7777778	Blue
9	468.7777778	476.25	Blue
10	476.25	483.7222222	Blue
11	483.7222222	491.1944444	Blue
12	491.1944444	498.6666667	Blue
13	498.6666667	506.1388889	Blue - Cyan
14	506.1388889	513.6111111	Cyan
15	513.6111111	521.0833333	Cyan - Green
16	521.0833333	528.5555556	Green
17	528.5555556	536.0277778	Green
18	536.0277778	543.5	Green
19	543.5	550.9722222	Green
20	550.9722222	558.4444444	Green
21	558.4444444	565.9166667	Green - Yellow
22	565.9166667	573.3888889	Yellow
23	573.3888889	580.8611111	Yellow
24	580.8611111	588.3333333	Yellow
25	588.3333333	595.8055556	Yellow - Orange
26	595.8055556	603.2777778	Orange
27	603.2777778	610.75	Orange
28	610.75	618.2222222	Orange
29	618.2222222	625.6944444	Orange - Red
30	625.6944444	633.1666667	Red
31	633.1666667	640.6388889	Red
32	640.6388889	648.1111111	Red

Tableau VII (suite)

Band Number b_i	Wavelength Range (nm)		Colour
33	648.1111111	655.5833333	Red
34	655.5833333	663.0555556	Red
35	663.0555556	670.5277778	Red
36	670.5277778	678	Red
37	678	685.4722222	Red
38	685.4722222	692.9444444	Red
39	692.9444444	700.4166667	Red
40	700.4166667	707.8888889	Red
41	707.8888889	715.3611111	Red
42	715.3611111	722.8333333	Red
43	722.8333333	730.3055556	Red
44	730.3055556	737.7777778	Red
45	737.7777778	745.25	Red - IR
46	745.25	752.7222222	IR
47	752.7222222	760.1944444	IR
48	760.1944444	767.6666667	IR
49	767.6666667	775.1388889	IR
50	775.1388889	782.6111111	IR
51	782.6111111	790.0833333	IR
52	790.0833333	797.5555556	IR
53	797.5555556	805.0277778	IR
54	805.0277778	812.5	IR
55	812.5	819.9722222	IR
56	819.9722222	827.4444444	IR
57	827.4444444	834.9166667	IR
58	834.9166667	842.3888889	IR
59	842.3888889	849.8611111	IR
60	849.8611111	857.3333333	IR
61	857.3333333	864.8055556	IR
62	864.8055556	872.2777778	IR
63	872.2777778	879.75	IR
64	879.75	887.2222222	IR
65	887.2222222	894.6944444	IR
66	894.6944444	902.1666667	IR
67	902.1666667	909.6388889	IR
68	909.6388889	917.1111111	IR
69	917.1111111	924.5833333	IR
70	924.5833333	932.0555556	IR
71	932.0555556	939.5277778	IR
72	939.5277778	947	IR

BIBLIOGRAPHIE

[1] H. O. Sparrow, "Nos sols dégradés. Rapport sur la conservation des sols au Sénat du Canada," Comité sénatorial permanent de l'agriculture, des pêches et des forêts, Canada, Hull (Québec, Canada) 6 novembre 1984.

[2] N. Zhang, M. Wang, and N. Wang, "Precision Agriculture--a worldwide overview," *Computers and electronics in agriculture*, vol. 36, pp. 113-132, 2002.

[3] Penuelas and Filella, "Technical Focus: Visible and near-infrared reflectance techniques for diagnosing plant physiological status.," *Trends Plant Sci.*, vol. 3, pp. 151-156, 1998.

[4] P. S. Thenkabail, R. B. Smith, and E. D. Pauw, "Hyperspectral Vegetation Indices and Their Relationships with Agricultural Crop Characteristics," *Remote Sensing of Environment*, vol. 71, pp. 158-182, 1999.

[5] D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, "*Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture*," *Remote Sensing of Environment*, vol. 81, pp. 416-426, 2002.

[6] N. H. Broge and E. Leblanc, "Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density," *Remote Sensing of Environment*, vol. 76, pp. 156-172, 2000.

[7] M. S. Moran, Y. Inoue, and E. M. Barnes, "Opportunities and limitations for image-based remote sensing in precision crop management," *Remote Sensing of Environment*, vol. 61, pp. 319-346, 1997.

[8] F. Baret, "Potentiel de la télédétection pour l'agriculture de précision," *L'enjeu Français de l'agriculture de précision*, pp. 8-11, 1999.

[9] D. S. Long, G. R. Carlson, and S. D. DeGloria, "Quality of field management maps," presented at Site-Specific Management for Agricultural Systems, Minneapolis, pp. 251-271, 1995.

[10] D. R. Nielsen, O. Wendroth, and M. B. Parlange, "Opportunities for examining on-farm soil variability," presented at Site-Specific Management for Agricultural Systems, Minneapolis, pp. 87-99, 1994.

[11] P. K. Goel, S. O. Prasher, J. A. Landry, R. M. Patel, R. B. Bonnel, A. A. Viau, and J. R. Miller, "Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn," *Computers and electronics in agriculture*, vol. 38, pp. 99-124, 2003.

[12] L. D. Hanson, P. C. Robert, and M. Bauer, "Mapping wild oats infestations using digital imagery for site-specific management," presented at Site-Specific Management for Agricultural Systems, Minneapolis, pp. 227-230, 1994.

[13] D. E. McGrath, A. V. Skotnikov, and V. A. Bobrov, "A site-specific expert system with supporting equipment for crop management," presented at Site-Specific Management for Agricultural Systems, Minneapolis, pp. 619-635, 1994.

[14] S. Jacquemoud, C. Bacour, H. Poilve, and J.-P. Frangi, "Comparison of Four Radiative Transfer Models to Simulate Plant Canopies Reflectance: Direct and Inverse Mode," *Remote Sensing of Environment*, vol. 74, pp. 471-481, 2000.

[15] W. Verhoef, "Light scattering by leaf layers with application to canopy reflectance modeling: the SAIL model," *Remote Sensing of Environment*, vol. 16, pp. 125-141, 1984.

[16] S. Jacquemoud and F. Baret, "PROSPECT: a model of leaf optical properties spectra," *Remote Sensing of Environment*, vol. 34, pp. 75-91, 1990.

[17] J. G. P. W. Clevers and W. Verhoef, "Modelling and synergetic use of optical and microwave remote sensing. LAI estimation from canopy reflectance and WDVI: a sensitivity analysis with the SAIL model,"in Report 90-39 of the Netherlands Remote Sensing Board (BCRS), 70 pp, 1991.

[18] S. W. Maier, W. Ludeker, and K. P. Gunther, "SLOP: A Revised Version of the Stochastic Model for Leaf Optical Properties," *Remote Sensing of Environment*, vol. 68, pp. 273-280, 1999.

[19] A. Kuusk, "A markov chain model of canopy reflectance model," *Agriculture for Meteorology*, vol. 76, pp. 221-236, 1995.

[20] S. Jacquemoud, F. Baret, B. Andrieu, F. M. Danson, and K. Jaggard, "Extraction of vegetation biophysical parameters by inversion of the PROSPECT+SAIL models on sugar beet canopy reflectance data. Application to TM and AVIRIS sensors," *Remote Sensing of Environment*, vol. 52, pp. 163-172, 1995.

[21] Peñuelas and Filella, "Technical Focus: Visible and near-infrared reflectance techniques for diagnosing plant physiological status.," *Trends in plant science*, vol. 3, pp. 151-156, 1998.

[22] S. Jacquemoud, "Utilisation de la haute résolution spectrale pour l'étude des couverts végétaux: développement d'un modèle de réflectance spectrale," in *Méthodes Physiques en Télédétection*. Paris: Paris VII, 1992, pp. 92.

[23] Y. J. Kaufman and D. Tanre, "Atmospherically resistant vegetation index (ARVI), for EOS-MODIS," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, pp. 261-270, 1992.

[24] A. R. Huete, C. Justice, and W. van Leeuwen, "Modis Vegetation index (MOD 13), EOS MODIS Algorithm -- Theoretical Basis Document," NASA Goddard Space Flight Ceter, Greenbelt, Maryland 20771, USA 1996.

[25] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crops canopies: Modeling and validation in the context of precision agriculture," *Remote Sensing of Environment*, vol. 90, pp. 337-352, 2004.

[26] C. S. T. Daughtry, C. L. Walthall, M. S. Kim, E. Brown de Colstoun, and J. E. McMurtrey, III, "Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance," *Remote Sensing of Environment*, vol. 74, pp. 229-239, 2000.

[27] J. A. Gamon and C. B. Field, "Relationship between NDVI, canopy structure and photosynthesis in three Californian vegetation types," *Ecological Applications*, vol. 5, pp. 28-41, 1995.

[28] J. Peñuelas, "Reflectance indices associated with physiological changes in nitrogen and water-limited sunflower leaves," *Remote Sensing of Environment*, vol. 48, pp. 135-146, 1994.

[29] J. Peñuelas, J. Piñol, R. Ogayar, and I. Fillela, "Estimation of plant water content by the reflectance Water Index WI (R900/R970)," *International Journal of Remote Sensing*, vol. 18, pp. 2869-2875, 1997.

[30] J. D. Barnes, "A reappraisal of the use of DMSO for the extraction and determination of chlorophylls a and b in lichens and higher plants," *Environmental Experimental Botany*, vol. 2, pp. 85-100, 1992.

[31] J. Peñuelas, "Reflectance assessment of plant mite attack on apple trees," *International Journal of Remote Sensing*, vol. 16, pp. 2727-2733, 1995.

[32] P. M. Hansen and J. K. Schjoerring, "Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indicies and partial least squares regression," *Remote Sensing of Environment*, vol. 86, pp. 542-553, 2003.

[33] D. Haboudane, J. R. Miller, N. Tremblay, E. Pattey, and P. Vigneault, "Estimation of leaf area index using ground spectral measurements over agricultural crops: prediction capability assessment of optical indices," presented at ISPRS 2004, Istanbul, pp. 108-113, 2004.

[34] J. L. Rougean and F. M. Breon, "Estimating PAR absorbed by vegetation bidirectional reflectance measurements," *Remote Sensing of Environment*, vol. 51, pp. 375-384, 1995.

[35] J. Chen, "Evaluation of vegetation indices and modified simple ratio for boreal applications," *Canadian Journal of Remote Sensing*, vol. 22, pp. 229-242, 1996.

[36] I. Filella, L. Serrano, J. Serra, and J. Peñuelas, "Evaluating wheat nitrogen status with canopy reflectance indices and discriminant analysis," *Crop science*, vol. 35, pp. 1400-1405, 1995.

[37] T. P. Dawson and P. J. Curran, "Technical Note: A new technique for interpolating the reflectance red edge position," *International Journal of Remote Sensing*, vol. 19, pp. 2133-2139, 1998.

[38] M. S. Kim, C. S. T. Daughtry, E. W. Chappelle, J. E. McMurtrey, and C. L. Walthall, "The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (Apar)," presented at Proceeding of the 6th Symp. on Physical Measurements and Signatures in Remote Sensing, Val D'Isère, France, pp. 299-306, 1994.

[39] H. K. Lichtenthaler and J. A. Miehé, "Fluorescence imaging as a diagnostic tool for plant stress," *Trends in plant science*, vol. 8, pp. 316-320, 1997.

[40] P. J. Zarco-Tejeda, J. R. Miller, G. H. Mohammed, T. L. Noland, and P. H. Sampson, "Chlorophyll fluorescence effects on vegetation apparent reflectance: II. Laboratory and Airborne Canopy-Level Measurements with Hyperspectral Data," *Remote Sensing of Environment*, vol. 74, pp. 596-608, 2000.

[41] P. J. Zarco-Tejada, J. R. Miller, G. H. Mohammed, and T. L. Noland, "Chlorophyll Fluorescence Effects on Vegetation Apparent Reflectance: I. Leaf-Level Measurements and Model Simulation," *Remote Sensing of Environment*, vol. 74, pp. 582-595, 2000.

[42] A. R. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sensing of Environment*, vol. 25, pp. 295-309, 1988.

[43] J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian, "A modified soil adjusted vegetation index," *Remote Sensing of Environment*, vol. 48, pp. 119-126, 1994.

[44] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. L. Marshak, "The Interpretation of Spectral Vegetation Indexes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, NO. 2, pp. 481-486, 1995.

[45] A. Bannari, "La télédétection et les indices de végétation," thesis in *Département de Géographie et de Télédétection.* Sherbrooke: Université de Sherbrooke, 1996.

[46] J. R. Koza, *Genetic Programming-On the programming of Computers by Means of Natural Selection.* Cambridge, London: MIT Press, 1992.

[47] J. Holland, *Adaptation in Natural and Artificial Systems*, 1 ed: University of Michigan Press (2nd ed in 1992, MIT Press), 1975.

[48] J. Holland, *Adaptation in Natural and Artificial Systems*, 2 ed. Cambridge, Massachusetts: MIT Press, 1992.

[49] C. Darwin, *On the Origin of Species by Means of Natural Selection.* Londres, 1859.

[50] R. Gross, K. Albrecht, W. Kantschik, and W. Banzhaf, "Evolving chess playing programs," presented at GECCO, pp. 740-747, 2002.

[51] H. Iba and E. Sakamoto, "Inference of Differential Equation Models by Genetic Programming," presented at GECCO, pp. 788-795, 2002.

[52] J. R. Koza, D. Andre, H. Forrest, I. Bennett, and M. A. Keane, "Use of automatically defined functions and architecture-altering operations in automated circuit synthesis using genetic programming," presented at Genetic Programming 1996: Proceedings of the First Annual Conference, Stanford University, Cambridge, MA, pp. 132-140, 1996.

[53] J. R. Koza, "Evolution of a computer program for classifying protein segments as transmembrane domains using genetic programming," presented at Intelligent Systems for Molecular Biology, pp. 244-252, 1994.

[54] S.-H. Chen and C.-H. Yeh, "On the emergent properties of artificial stock markets: the efficient market hypothesis and the rational expectations

hypothesis," *Journal of Economic Behavior & Organization*, vol. 49, pp. 217-239, 2002.

[55] S. P. Brumby, P. A. Pope, A. E. Galbraith, and J. J. Szyinanski, "Evolving feature extraction algorithms for hyperspectral and fused imagery," presented at Proceedings of Fifth International Conference on Information Fusion, 8-11 July 2002, Annapolis, MD, USA, pp. 986-993, 2002.

[56] B. J. Ross, A. G. Gualtieri, F. Fueten, and P. Budkewitsch, "Hyperspectral Image Analysis Using Genetic Programming," presented at Genetic and Evolutionary Computation Conference (GECCO-2002), pp. 1196-1203, 2002.

[57] M. J. Aitkenhead, I. A. Dalgetty, C. E. Mullins, A. J. S. McDonald, and I. B. Strachan, "Weed and crop discrimination using image analysis and artificial intelligence methods," *Computers and Electronics in Agriculture*, vol. 39, pp. 157-171, 2003.

[58] MengBo Li and R. S. Yost, "Management-oriented modeling: optimizing nitrogen management with artificial intelligence," *Agricultural Systems*, vol. 65, pp. 1-27, 2000.

[59] L. M. Dwyer, A. M. Anderson, B. L. Ma, D. W. Stewart, M. Tollenaar, and E. Gregorich, "Quantifying the nonlinearity in chlorophyll meter response to corn leaf nitrogen concentration," *Canadian Journal of plant science*, vol. 75, pp. 179-182, 1994.

[60] C. Cowger and C. Mundt, "A hydroponic seedling assay for resistance to cephalosporium stripe of wheat," *Plant disease*, vol. 82, pp. 1126-1131, 1998.

[61] B. B. Mehdi, C. A. Madramootoo, and G. R. Mehuys, "Yield and Nitrogen of Corn under Different Tillage Practices," *Agronomy Journal*, vol. 91, pp. 631-636, 1999.

[62] I. B. Strachan, E. Pattey, and J. B. Boisvert, "Impact of nitrogen environmental conditions on corn as detected by hyperspectral reflectance," *Remote Sensing of Environment*, vol. 80, pp. 213-224, 2002.

[63] B. O. Hoel and K. A. Solhaug, "Effect of irradiance on chlorophyll estimation with the Minolta SPAD-502 leaf chlorophyll meter," *Annals of Botany*, vol. 82, pp. 389-392, 1998.

[64] D. E. Martínez and J. J. Guiamet, "Distortion of the SPAD 502 chlorophyll meter readings by changes in irradiance and leaf water status," *Agronomie*, vol. 24, pp. 41-46, 2004.

[65] W. B. Langdon, "Size Fair and Homologous Tree Crossovers for Tree Genetic Programming," *Genetic Programming and Evolvable Machines*, vol. 1, pp. 95-119, 2000.

[66] E. De Jong and J. Pollack, "Multi-Objective Methods for Tree Size Control," *Genetic Programming and Evolvable Machines*, vol. 4, pp. 211-233, 2003.

[67] R. Crawford-Marks and L. Spector, "Size Control via Size Fair Genetic Operators in the PushGP Genetic Programming System," presented at Genetic and Evolutionary Computation Conference, San Francisco, pp. 733-739, 2002.

[68] S. Luke and L. Panait, "Fighting Bloat With Nonparametric Parsimony pressure," presented at 7th International Conference on Parallel Problem Solving from Nature, pp. 411-421, 2002.

[69] Peter W.H. Smith, "Controlling Code Growth in Genetic Programming," in *Soft computing techniques and applications*, R. John and R. Birkenhead, Eds., 2000, pp. 166-171.

[70] S. l. Osborne, J. S. Shepers, D. D. Francis, and M. R. Schlemmer, "Detection of Phosphorus and Nitrogen Deficiencies in Corn Using Spectral Radiance Measurements," *Agronomy Journal*, vol. 94, pp. 1215-1221, 2002.

[71] S. Kumar, J. Ghosh, M. M. Crawford, and I. Member, "Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1368-1379, 2001.

[72] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote sensing image display and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 538-542, 1999.

[73] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2653-2667, 1999.

[74] M.-C. Girard and C. M. Girard, *Traitement des données de télédétection*. Paris, 1999.

[75] T. M. Lillesand and R. W. Kiefer, *Remote sensing and image interpretation*, fourth ed, 2000.