



# THÈSE DE DOCTORAT DE L'UNIVERSITÉ MICHEL DE MONTAIGNE - BORDEAUX 3

École Doctorale Montaigne Humanités (ED 480)

**Discipline: Science et Technologie**  
**Spécialité: Informatique**

présentée par

Benoît BEGUET

pour obtenir le grade de:

Docteur de l'Université Michel de Montaigne

---

**Caractérisation et cartographie de la structure forestière à partir d'images satellitaires à très haute résolution spatiale.**

---

Directrice de thèse: Pr Samia BOUKIR

Soutenue le 06 10 2014 devant le jury composé de:

M. Jordi INGLADA	HDR, Chercheur CESBIO	Président du Jury
M. Nicolas PAPANODITIS	HDR, DR IGN	Rapporteur
M. Pierre COUTERON	HDR, IRD UMR AMAP	Rapporteur
Mme. Samia BOUKIR	Pr, ENSEGID-IPB	Directrice de thèse
Mme. Dominique GUYON	IR, INRA	Co-Encadrante
Mme. Nesrine CHEHATA	MCU, ENSEGID-IPB	Co-Encadrante
Mme. Anne JOLLY	IR, ONF	Invitée



## Résumé

Les images à très haute résolution spatiale (THR) telles que les images Pléiades (50 cm en Panchromatique, 2m en multispectral) rendent possible une description fine de la structure forestière (distribution et dimensions des arbres) à l'échelle du peuplement, en exploitant la relation entre la structure spatiale des arbres et la texture d'image quand la taille du pixel est inférieure à la dimension des arbres. Cette attente répond au besoin d'inventaire spatialisé de la ressource forestière à l'échelle du peuplement et de ses changements dus à la gestion forestière, à l'aménagement du territoire ou aux événements catastrophiques. L'objectif est double: (1) évaluer le potentiel de la texture d'images THR pour estimer les principales variables de structure forestière (diamètre des couronnes, diamètre du tronc, hauteur, densité ou espacement des arbres) à l'échelle du peuplement; (2) sur ces bases, classer les données image, au niveau pixel, par types de structure forestière afin de produire l'information spatialisée la plus fine possible. Les principaux développements portent sur l'automatisation du paramétrage, la sélection de variables, la modélisation par régression multivariable et une approche de classification par classifieurs d'ensemble (Forêts Aléatoires ou Random Forests). Ils sont testés et évalués sur deux sites de la forêt landaise de pin maritime à partir de trois images Pléiades et une Quickbird, acquises dans diverses conditions (saison, position du soleil, angles de visée). La méthodologie proposée est générique. La robustesse aux conditions d'acquisition des images est évaluée. Les résultats montrent que des variations fines de texture caractéristiques de celles de la structure forestière sont bien identifiables. Les performances en terme d'estimation des variables forestières (RMSE) : 1.1 m pour le diamètre des couronnes, 3 m pour la hauteur des arbres ou encore 0.9 m pour leur espacement, ainsi qu'en cartographie des structures forestières (82% de taux de bonne classification pour la reconnaissance des 5 classes principales de la structure forestière) sont satisfaisantes d'un point de vue opérationnel. L'application à des images multi-annuelles permettra d'évaluer leur capacité à détecter et cartographier des changements tels que coupe forestière, mitage urbain ou encore dégâts de tempête.



## Abstract

Very High spatial Resolution (VHR) images like Pléiades imagery (50 cm panchromatic, 2m multispectral) allows a detailed description of forest structure (tree distribution and size) at stand level, by exploiting the spatial relationship between tree structure and image texture when the pixel size is smaller than tree dimensions. This information meets the expected strong need for spatial inventory of forest resources at the stand level and its changes due to forest management, land use or catastrophic events. The aim is twofold : (1) assess the VHR satellite images potential to estimate the main variables of forest structure from the image texture: crown diameter, stem diameter, height, density or tree spacing, (2) on these bases, a pixel-based image classification of forest structure is processed in order to produce the finest possible spatial information. The main developments concern parameter optimization, variable selection, multivariate regression modelling and ensemble-based classification (Random Forests). They are tested and evaluated on the Landes maritime pine forest with three Pléiades images and a Quickbird image acquired under different conditions (season, sun angle, view angle). The method is generic. The robustness of the proposed method to image acquisition parameters is evaluated. Results show that fine variations of texture characteristics related to those of forest structure are clearly identifiable. Performances in terms of forest variable estimation (RMSE): 1,1m for crown diameter, 3m for tree height and 0,9m for tree spacing, as well as forest structure mapping (82% Overall accuracy for the classification of the five main forest structure classes) are satisfactory from an operational perspective. Their application to multi-annual images will assess their ability to detect and map forest changes such as clear cut, urban sprawl or storm damages.



---

## Avant Propos

Ce travail a été réalisé à l'institut ENSEGID-IPB (Ecole Nationale Supérieure Environnement Géoressource et Ingénierie du Développement durable, Institut Polytechnique de Bordeaux) à Pessac, ainsi qu'à l'Unité de Recherche ISPA (Interaction Sol Plante Atmosphère), anciennement EPHYSE (Ecologie Fonctionnelle et Physique de l'Environnement) de l'INRA (Institut National de la Recherche Agronomique) à Villenave d'Ornon. La thèse a été financée par le CNES (Centre National des Etudes Spatiales) à Toulouse ainsi que par la Région Aquitaine. La thèse a été menée sous la direction de Mme Samia Boukir avec l'encadrement scientifique de Mme Dominique Guyon et Mme Nesrine Chehata.



# Table des matières

Avant Propos . . . . .	v
<b>Table des figures</b>	<b>xi</b>
<b>Liste des tableaux</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Contexte . . . . .	2
1.1.1 La forêt landaise . . . . .	2
1.1.2 Apport des images THR pour les applications forestières . . . . .	3
1.2 Objectifs de la thèse . . . . .	5
<b>2 Bases bibliographiques et démarche adoptée</b>	<b>7</b>
2.1 Texture : définition . . . . .	7
2.2 Texture THR et forêts . . . . .	8
2.3 Analyse de texture . . . . .	11
2.3.1 Statistiques de texture du premier ordre . . . . .	11
2.3.2 Semivariogrammes . . . . .	12
2.3.3 Matrice de co-occurrence (Grey Level Cooccurrence Matrix ou GLCM) . . . . .	13
2.3.4 Approches fréquentielles . . . . .	16
2.3.5 Autres approches . . . . .	18
2.4 Démarche adoptée . . . . .	19
<b>3 Données</b>	<b>23</b>
3.1 Sites d'étude . . . . .	23
3.2 Données images . . . . .	24
3.3 Informations forestières de référence . . . . .	27
3.3.1 Mesures dendrométriques in situ sur les deux sites . . . . .	27
3.3.2 Photo-interprétation sur le site du Tagon . . . . .	30

<b>4</b>	<b>Estimation des variables forestières à partir de la texture THR par modélisation statistique</b>	<b>35</b>
4.1	Introduction . . . . .	35
4.2	Méthodes de sélection de variables par régression linéaire . . . . .	36
4.2.1	Détection de la colinéarité . . . . .	37
4.2.2	Méthodes de sélection de variables . . . . .	39
4.3	Méthodologie . . . . .	42
4.3.1	Démarche proposée . . . . .	42
4.3.2	Attributs de texture et paramètres . . . . .	43
4.3.3	Échantillons d'image . . . . .	45
4.3.4	Sélection de variables par régressions linéaires . . . . .	47
4.4	Résultats . . . . .	49
4.4.1	Choix préliminaires . . . . .	49
4.4.2	Solutions par régression simple . . . . .	50
4.4.3	Solutions par régression multiple . . . . .	60
4.5	Conclusion . . . . .	81
<b>5</b>	<b>Cartographie des types de structures forestières</b>	<b>83</b>
5.1	Introduction . . . . .	83
5.2	Méthodes de sélection de variables pour la classification . . . . .	85
5.2.1	Introduction . . . . .	85
5.2.2	Bases théoriques, taxonomie des méthodes de sélection de variables : . . . . .	86
5.2.3	Applications en classification d'images de télédétection . . . . .	90
5.3	Les forêts aléatoires : classifieur ensembliste . . . . .	91
5.3.1	Algorithme . . . . .	92
5.3.2	Erreur OOB . . . . .	93
5.3.3	Marge d'ensemble . . . . .	94
5.4	Méthodologie : vers une stratégie opérationnelle . . . . .	95
5.5	Recherche d'une méthode de sélection d'attributs optimale . . . . .	98
5.5.1	Expérience sur le site de Nezer . . . . .	98
5.5.2	Expérience complémentaire sur le site du Tagon . . . . .	103
5.6	Cartographie des structures forestières : vers une solution opérationnelle . . . . .	106
5.6.1	Une approche hiérarchique . . . . .	106
5.6.2	Utilisation de la marge d'ensemble : . . . . .	108
5.6.3	Evaluation multi-dates . . . . .	118
5.7	Conclusions . . . . .	125

---

<b>6</b>	<b>Conclusions et perspectives</b>	<b>127</b>
6.1	Analyse de texture . . . . .	128
6.2	Estimation des variables forestières . . . . .	128
6.3	Cartographie des structures forestières . . . . .	129
6.4	Perspectives . . . . .	129
6.5	Conclusion générale . . . . .	131
	<b>Références</b>	<b>133</b>
	<b>Annexes</b>	<b>145</b>
<b>A</b>	<b>Protocole de mesures au sol de variables de structure forestière sur un échantillon de peuplements de pin maritime sur le site Tagon-Marcheprime.</b>	<b>147</b>
<b>B</b>	<b>Analyse comparative des méthodes de sélection d'attributs : tableaux de résultats complémentaires</b>	<b>155</b>
<b>C</b>	<b>Retrieving forest structure variables from very high resolution satellite images using an automatic method.</b>	<b>163</b>
<b>D</b>	<b>Modelling-based feature selection for classification of forest structure using very high resolution multispectral imagery.</b>	<b>171</b>
<b>E</b>	<b>Automated retrieval of forest structure variables based on multi-scale texture analysis of VHR satellite imagery.</b>	<b>179</b>
<b>F</b>	<b>Classification of forest structure using very high resolution Pléiades image texture.</b>	<b>195</b>
<b>G</b>	<b>Quantification et cartographie de la structure forestière à partir de la texture des images Pléiades.</b>	<b>201</b>



# Table des figures

1.1	Présentation des différentes applications de la télédétection en milieu forestier en fonction de la résolution spatiale. Les croix entre parenthèses représentent les application visées par les images THR. D'après : Jolly (2014). . . . .	4
2.1	Représentation schématique de la construction d'une GLCM. Les paramètres sont $r$ pour le rayon de la fenêtre de calcul, $xd$ et $yd$ pour la longueur (et l'orientation) du déplacement, $nbbin$ pour le nombre de niveaux de quantification. Est représenté en rouge le comptage d'une paire de niveau de quantification 4 dans la fenêtre de calcul (en vert) selon le déplacement et l'orientation définis dans cet exemple. . . . .	13
3.1	Localisation des sites d'étude. Forêt des Landes de Gascogne (triangle vert), Aquitaine, France. . . . .	23
3.2	Site de Nezer, localisation des peuplements et des placettes de mesure (n=12, points verts). Image de fond Quickbird panchromatique. . . . .	25
3.3	Site de Tagon, localisation des peuplements de pin étudiés et des placettes de mesure (n=111, points verts). Image de fond Pléiades pancromatique du mois de juin. . . . .	26
3.4	Carte des âges des peuplements sur le site de Nezer en 2003. . . . .	28
3.5	Variables forestières mesurées sur les deux sites : diamètre moyen des couronnes ( $Cd$ ), hauteur moyenne des arbres ( $Ht$ ), espacement entre les arbres ( $Sp$ ) et densité des tiges ( $Nah$ ) en fonction du diamètre des troncs moyen ( $Dbh$ ). . . . .	31

3.6	Variables forestières mesurées et classes d'appartenance : diamètre des couronnes ( $Cd$ ), hauteur des arbres ( $Ht$ ) et espacement entre les arbres ( $Sp$ ) en fonction du diamètre des troncs ( $Dbh$ ). C2 à C6 : classes de structure à priori (cf tableau 3.3) c'est à dire données par photo-interprétation avant la campagne de terrain. Noter l'erreur d'affectation sur 2 peuplements (C4 au lieu de C3) corrigée par la suite. . . . .	33
3.7	Imagettes panchromatiques et classes de la structure forestière associée. C1 = pas d'arbres visibles (coupes rases, reboisement en cours ou récent), âge des peuplements $\leq 3$ ans; C2 = arbres visibles, rangs visibles, $Cd < 1m$ ; C3 = $Cd \in [1m;2.5m]$ , rangs visibles; C4 = $Cd \in [2.5m;4.5m]$ , rangs rarement visibles, densité plus faible; C5 = $Cd \geq 4.5m$ , rangs non visibles, densité faible. . . . .	34
4.1	Représentation schématique générale de la modélisation des variables forestières à partir des attributs image. . . . .	43
4.2	Echantillons d'images panchromatiques pour les deux sites illustrant la variabilité des peuplements en termes de structure forestière. Les échantillons sont ordonnés par âge croissant. . . . .	46
4.3	Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site de Nezer. Les barres correspondent à $\pm 1$ écart type. . . . .	56
4.4	Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Juin. Les barres correspondent à $\pm 1$ écart type. . . . .	57
4.5	variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Aout. Les barres correspondent à $\pm 1$ écart type. . . . .	58
4.6	Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Février. Les barres correspondent à $\pm 1$ écart type. . . . .	59
4.7	Evolution de la statistique $C_p$ de Mallow en fonction du nombre d'attributs de texture pour l'estimation du diamètre des couronnes et de l'âge des peuplements pour le site de Nezer (image d'octobre). . . . .	60
4.8	Evolution de la statistique $C_p$ de Mallow en fonction du nombre d'attributs de texture pour l'estimation du diamètre des couronnes pour le site du Tagon en Juin et Aout à partir des trois ensembles de données MS, Pan et MS-Pan. . . . .	61

4.9	Estimation de l'âge des peuplements, LARS stepwise, site de Nezer (n=184). LOO RMSE = 5.25 ans. . . . .	64
4.10	Variables de structure forestière mesurées vs. estimations prédites par des meilleurs modèles multiples, site de Nezer. Les barres correspondent à $\pm 1$ écart type. . . . .	73
4.11	variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Juin. Les barres correspondent à $\pm 1$ écart type. . . . .	74
4.12	variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Aout. Les barres correspondent à $\pm 1$ écart type. . . . .	75
4.13	variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Février. Les barres correspondent à $\pm 1$ écart type. . . . .	76
4.14	Variation de la RMSE moyenne en fonction du nombre d'échantillons d'apprentissage. Résultats pour <i>Cd</i> , <i>Ht</i> et <i>Sp</i> avec le même échantillon de test (n=40) et 10 répétitions par taille d'échantillon d'apprentissage. Les barres représentent $\pm 1$ écart type. Site du Tagon, images PLéiades. . . . .	80
5.1	Représentation schématique de la stratégie opérationnelle globale envisagée. . . . .	96
5.2	Image multispectrale <i>Quickbird</i> (ROI 2), $1136 \times 972$ pixels, forêt de Nezer (composition colorée PIR, R, V). . . . .	99
5.3	Classification de l'image multispectrale <i>Quickbird</i> (ROI 2) du site de <i>Nezer</i> utilisant la sélection d'attributs par modélisation . . . .	102
5.4	Observation de l'évolution de la précision globale de classification en fonction du nombre d'attributs de texture, image de juin. . . .	106
5.5	Classification C1 versus C2 (Autres) et carte de marge associée, image de juin. . . . .	111
5.6	Classification C2 à C5 et carte de marge associée, image de juin. .	112
5.7	Echantillons de référence après corrections d'étiquetage, fond : image panchromatique du mois de juin. . . . .	114
5.8	Exemple d'une parcelle complexe et de certaines difficultés rencontrées par le classifieur, image Pléiades du mois de juin. . . . .	116
5.9	Exemple de localisation d'une trouée dans un peuplement mature (classe 5), image Pléiades du mois de juin. . . . .	117
5.10	Classification finale, mois de juin. . . . .	120

5.11 Classification finale, mois d’Août. . . . .	122
5.12 Classification finale, mois de Février. . . . .	124

# Liste des tableaux

2.1	Descriptions des attributs de texture issus des GLCM utilisés dans cette étude. . . . .	16
3.1	Paramètres d'acquisition des images. . . . .	27
3.2	Gamme de variation des variables forestières étudiées (Nezer : 12 peuplements, Tagon : 111 peuplements). Pour Cd, Ht et Dbh, les valeurs moyennes sont présentées. . . . .	30
3.3	Descriptif des classes de structures forestières. . . . .	32
4.1	Gammes des valeurs des paramètres des GLCM. . . . .	44
4.2	Les attributs de texture I, MS et Pan les plus performants pour l'estimation des variables forestières. Site de Nezer, image Quick-bird d'Octobre. . . . .	50
4.3	Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Nezer, image d'Octobre, n=12. . . . .	51
4.4	Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image de Juin, n=111. . . . .	52
4.5	Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image d'Août, n=111. . . . .	52
4.6	Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image de Février, n=111. . . . .	53
4.7	Comparaison des performances des méthodes de sélection de variables pour l'estimation de l'âge, données panchromatiques ( $n = 12$ ). Nezer . . . . .	63

---

4.8	Comparaison des performances des méthodes de sélection de variables pour l'estimation de l'âge, données multispectrales ( $n = 12$ ). Nezer . . . . .	63
4.9	Analyse comparative des méthodes de sélection d'attributs : critères statistiques des meilleurs sous ensembles de trois attributs (MS-Pan), image d'Octobre, site de Nezer. . . . .	66
4.10	Analyse comparative des méthodes de sélection d'attributs : critères statistiques des meilleurs sous ensembles de cinq attributs (Pan), image d'Août, site de Tagon. . . . .	66
4.11	Importance de variables pour les sous-ensembles d'attributs sélectionnés (MS-Pan) par LARS stepwise et Random PRiF, pour les sites de Nezer (image d'Octobre) et du Tagon (image d'Août). . .	67
4.12	Meilleures solutions multiples pour les sites de Nezer et du Tagon (Août) avec la méthode de sélection d'attribut retenue et l'ensemble de données respectif. . . . .	68
4.13	Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Nezer. . . . .	69
4.14	Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image de Juin. . . . .	70
4.15	Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image d'Août. . . . .	71
4.16	Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image de Février. . . . .	72
4.17	Comparaison des performances des solutions simples et multiples (MS-Pan) pour les deux sites. . . . .	77
5.1	Performances de classification globale (%) par différentes stratégies de sélection de variables sur deux échantillons de l'image multispectrale <i>Quickbird</i> , avec 3 attributs de texture. . . . .	101
5.2	Matrice de confusion pour la classification de l'échantillon ROI 2 par Bagging Trees utilisant les 3 meilleurs attributs de texture décrivant la variable Sp (espacement entre les arbres) . . . . .	101
5.3	Performances de classification globale (%) par différentes stratégies de sélection de variables sur deux échantillons de l'image multispectrale <i>Quickbird</i> , avec 15 attributs de texture . . . . .	103

---

5.4	Matrice de confusion obtenue avec les meilleurs attributs de texture (n=50) issus de la sélection de variables Random PRiF pour le site du Tagon, image de Juin. . . . .	104
5.5	Matrice de confusion obtenue avec les meilleurs attributs de texture (n=50) obtenus par importance de variables des forêts aléatoires (attribut par attribut et bande spectrale par bande spectrale), image de Juin. . . . .	104
5.6	Matrice de confusion obtenue avec les meilleurs attributs de texture (n=50) obtenus par importance de variables des forêts aléatoires (de façon globale) pour le site du Tagon, image de Juin. . . . .	105
5.7	Matrice de confusion obtenue par classification directe de C1 à C5, image de Juin, 15 attributs de texture. . . . .	107
5.8	Matrice de confusion obtenue C2-C5, image de Juin, 15 attributs de texture. . . . .	109
5.9	Matrice de confusion obtenue C1 VS autres image de Juin, 15 attributs de texture. . . . .	109
5.10	Résumé des alarmes et corrections d'étiquetage détectés avec un seuil de marge moyenne à -0.3. . . . .	113
5.11	Matrice de confusion obtenue pour C1 VS. Autres, image de Juin. . . . .	119
5.12	Matrice de confusion obtenue pour C2-C5, image de Juin. . . . .	119
5.13	Matrice de confusion obtenue pour C1-C5, image de Juin. . . . .	119
5.14	Matrice de confusion obtenue pour C1 Vs. Autres image d'Août. . . . .	121
5.15	Matrice de confusion obtenue pour C2-C5, image d'Août. . . . .	121
5.16	Matrice de confusion obtenue pour C1-C5, image d'Août. . . . .	121
5.17	Matrice de confusion obtenue pour C1 Vs. Autres image de février. . . . .	123
5.18	Matrice de confusion obtenue pour C2-C5, image de février. . . . .	123
5.19	Matrice de confusion obtenue pour C1-C5, image de février. . . . .	123
B.1	Feature selection methods comparison : performance of the best 3 features model using the MS dataset of october image on Nezer site.	156
B.2	Feature selection methods comparison : performance of the best 5 features model using the MS dataset of august image on Tagon site.	156
B.3	Feature selection methods comparison : performance of the best 3 features model using the Pan dataset of october image on Nezer site.	157
B.4	Feature selection methods comparison : performance of the best 5 features model using the Pan dataset of august image on Tagon site.	157

B.5	Feature selection methods comparison : performance of the best 3 features model using the MS-Pan dataset of october image on Nezer site. . . . .	158
B.6	Feature selection methods comparison : performance of the best 5 features model using the MS-Pan dataset of august image on Tagon site. . . . .	158
B.7	Comparison of subsets balance between LARS stepwise and Random PRiF, Nezer Site, MS dataset . . . . .	159
B.8	Comparison of subsets balance between LARS stepwise and Random PRiF, Nezer Site, Pan dataset . . . . .	159
B.9	Variable importance in subsets provided by LARS stepwise and Random PRiF feature selection methods, Nezer site, MS-PAN dataset. . . . .	160
B.10	Variable importance in subsets provided by LARS stepwise and Random PRiF feature selection methods, Tagon site (June), MS dataset. . . . .	160
B.11	Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). MS dataset . . . . .	161
B.12	Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). Pan dataset . . . . .	161
B.13	Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). MS-Pan dataset . . . . .	162

# Chapitre 1

## Introduction

L'étude et la surveillance des écosystèmes forestiers, dans un contexte de changements climatiques et de diminution de la biodiversité, est une priorité scientifique, politique et socio-économique. Ce suivi devant couvrir de grands espaces, la télédétection et plus particulièrement l'imagerie satellitaire représente une solution peu coûteuse et prometteuse. Depuis une dizaine d'années, le nombre de données satellitaires optiques à Très Haute Résolution (THR) spatiale ne cesse d'augmenter, fournissant des données de précision à divers utilisateurs. Dans le contexte du programme Orfeo (du CNES, préparatoire à l'utilisation des images Pléiades), cette étude a pour objectif d'explorer les capacités des images THR pour la caractérisation de la structure du couvert forestier. A l'échelle du peuplement, la forte relation entre la structure spatiale de la forêt et la texture présente dans l'image devrait fournir une information intéressante. Cette étude est focalisée sur la forêt de pin maritime des Landes de Gascogne et l'utilisation des images Pléiades.

Après avoir introduit le contexte de la forêt landaise et ses enjeux, nous présenterons le potentiel des images THR et les avancées opérationnelles attendues pour des applications forestières. Enfin nous préciserons les objectifs de cette thèse avant d'en annoncer le plan.

## 1.1 Contexte

### 1.1.1 La forêt landaise

Le massif des Landes de Gascogne, aussi nommé forêt des Landes de Gascogne, forêt landaise, massif gascon et massif landais, constitue le premier massif forestier français en superficie. Son périmètre englobe 1,33 million d'hectares et s'étend sur les départements de la Gironde, des Landes et du Lot-et-Garonne du sud-ouest de la France. La surface forestière de production recouvre environ 1 million d'hectares dont près de 90% est occupée par des peuplements de pin maritime (*Pinus Pinaster* Ait.). Ce massif est le premier massif forestier cultivé d'Europe. C'est le premier producteur de bois rond en France avec 23% de la production nationale toutes espèces confondues et 90% de la production nationale de pin maritime. Le pin maritime, dont l'installation généralisée date du début du 19<sup>ième</sup> siècle est traitée en futaie régulière et gérée de manière intensive. Le massif Landais représente donc un enjeu environnemental et socio-économique très important à l'échelle régionale mais aussi nationale (Yauschew-Raguenes, 2012).

Les principales menaces pour le massif landais et sa production sont les incendies, les tempêtes et les défoliations ou mortalités dues aux attaques d'insectes (chenille processionnaire du pin et scolytes principalement). Les changements climatiques impliquent une pression de plus en plus grande pour les forestiers qui doivent faire face à des catastrophes naturelles maintenant récurrentes comme les tempêtes Martin (1999) et Klaus (2009) qui à elle deux ont ravagé 50% du volume de bois de pin maritime sur pied. Le massif landais est donc une zone aussi importante que fragile. Aussi, la forêt étant un vaste domaine, il est très difficile de l'observer et d'observer ses évolutions de manière spatialisée dans un temps court (et notamment dans l'urgence consécutive aux catastrophes naturelles). Il apparaît alors nécessaire de développer des outils d'aide à l'observation, à l'estimation, au suivi de la ressource forestière, en se basant sur des données géographiques de précision comme les images satellitaires. Cette idée n'est pas nouvelle. Elle a déjà été explorée par le passé avec des images à haute résolution (10-30m, Spot et LandsatTM). Ainsi les travaux de Jolly (1993); Jolly et al. (1993, 1996) sont à l'origine de la cartographie opérationnelle des coupes rases annuelles et des dégâts des tempêtes Martin et Klaus sur l'ensemble du massif landais par le service de l'Inventaire Forestier National. L'avènement récent de la THR a ouvert de nouvelles perspectives sur le plan non seulement de la précision géométrique, mais aussi thématique.

### 1.1.2 Apport des images THR pour les applications forestières

Depuis une dizaine d'années, le nombre de capteur satellitaires fournissant des images à Très Haute Résolution spatiale (THR) ne cesse d'augmenter, les plus connus étant Ikonos, Quickbird, Geoeye et la nouvelle constellation de satellites Pléiades développée par le CNES. Les images THR permettent grâce à leurs résolutions sub-métriques de voir de nombreux détails dans l'image et fournissent aussi une texture très informative. De nombreuses représentations de la texture sont proposées dans la littérature et appliquées aux données de télédétection pour une vaste gamme d'applications comme par exemple la cartographie du milieu urbain (Dell'Acqua and Gamba, 2003; Pacifici et al., 2009; Soe and Tyler, 2004), la cartographie de la végétation et des habitats naturels (Tuttle et al., 2006; Wood et al., 2012) ou encore la segmentation des images (Gaetano et al., 2009; Trias-Sanz et al., 2008).

Si de nombreuses études ont déjà été menées sur l'étude du couvert forestier à partir d'images satellitaires avec des résolutions spatiales moindres (exemple Spot ou Landsat), l'utilisation d'images THR apporte une information nouvelle, beaucoup plus fine et détaillée, avec des pixels dont la taille est bien souvent inférieure à la dimension des arbres. Lorsque le pixel est plus grand que l'arbre (supérieur à 10m), la forêt si elle n'est pas très hétérogène spatialement (grandes trouées, agrégation des arbres par paquets de taille variable) apparaît "lisse", c'est à dire que la résolution ne permet pas de décrire la complexité de la structure du couvert forestier. Les indicateurs images caractérisent alors la forêt à l'échelle du peuplement en fonction de l'intensité du signal radiométrique. Lorsque le pixel est inférieur à la taille des arbres, l'image présente une texture avec des contrastes importants, on peut alors observer de nombreux détails invisibles jusqu'alors. La texture des image de forêts présente une grande complexité qui va de structures en rang pour les peuplements les plus jeunes jusqu'à des structures très hétérogènes pour les plus âgés. La complexité de la structure forestière, au sein d'un même peuplement, est observable. Les images THR permettent donc, à travers leur texture, d'observer différents types de structure forestière ainsi que des variations spatialement fines.

Cet apport principal des images THR pour l'étude de la forêt, la texture, permet d'imaginer différentes applications opérationnelles qui n'étaient pas envisageables avec des résolutions spatiales moindres. La figure 1.1 récapitule les différentes applications de la télédétection pour l'étude des forêts et le niveau de

résolution requis pour ces dernières.

**Résolution spatiale nécessaire :** **GROSSIERE**  **FINE**

	Niveau d'utilisation de l'information				
	Global	National/ Régional	Sub-Régional	Massif Forestier	Peuplement
<b>Suivi global</b>					
-Couvert forestier	X	X	X		
-Déforestation					
-Modélisation biomasse/carbone					
-Principales espèces et types de forêts					
<b>Catastrophes naturelles</b>					
-Incendies (détection, suivi de restauration)	X	X	X	X	
-Tempêtes			X	(X)	
<b>Déforestation, exploitation illégale</b>					
-Détection	X	X	X		
<b>Défoliations massives</b>					
-Détection des dommages et suivi			X	X	
<b>Inventaires forestiers et gestion</b>					
-Types forestiers et espèces			X	(X)	(X)
-Structure forestière : estimation de variables dendrométriques				(X)	(X)
-cartographie des habitats naturels				(X)	(X)

Figure 1.1 : Présentation des différentes applications de la télédétection en milieu forestier en fonction de la résolution spatiale. Les croix entre parenthèses représentent les applications visées par les images THR. D'après : Jolly (2014).

Les applications visées par l'utilisation d'images THR sont donc des applications qui nécessitent une information à l'échelle du peuplement, mais aussi à l'intérieur du peuplement afin d'appréhender la variabilité intra-parcellaire. Plus précisément, les méthodologies à développer autour des images THR doivent apporter des informations sur la structure du couvert forestier, notamment des informations quantitatives sur les variables caractéristiques de la structure forestière comme par exemple la hauteur des arbres ou le diamètre des couronnes qui permettent, via des modèles, d'estimer le volume du bois ou de la biomasse sur pied.

Des informations finement spatialisées en cas de catastrophe naturelle comme les tempêtes doivent permettre de localiser et d'appréhender l'importance des dégâts (même les plus diffus). Enfin d'autres attentes portent sur le suivi de la gestion forestière en particulier la détection des coupes (éclaircies, coupes rases), des reboisements et le suivi des régénérations.

## 1.2 Objectifs de la thèse

L'objectif principal de ce travail est d'explorer le potentiel des images THR pour la caractérisation du couvert forestier, en développant un cadre méthodologique tourné vers l'opérationnel. Les applications thématiques visées sont l'estimation des variables forestière d'intérêt (le diamètre des couronnes, la hauteur des arbres, l'espacement entre les arbres, le diamètre des troncs à hauteur de poitrine, la densité des tiges) et la cartographie des types de structure forestière, du stade de la forêt en mutation aux différents stades de maturité des peuplements (en vue d'inventaires forestiers finement spatialisés) à partir de la texture des images. Les attentes d'un point de vue méthodologique sont l'adaptation et l'optimisation de méthodes d'analyse d'images et de traitement de l'information existantes pour optimiser l'exploitation des informations fournies par les capteurs THR. Ces deux objectifs principaux, une fois atteints, formeront une base solide pour développer des applications de détection de changements et de suivi temporel (non abordées dans ce travail).

Dans un premier temps (Chapitre 2) un état de l'art sur les outils d'analyse de texture les plus rencontrés en télédétection est présenté avant de préciser nos choix et la démarche de travail adoptée. Le Chapitre 3 présente les sites d'étude et les données utilisées durant ce travail de thèse. Le Chapitre 4 porte sur l'estimation des variables forestières à partir de la texture des images THR. Le chapitre 5 présente la cartographie des types de structure forestière (classification d'images) à partir des images THR et enfin les conclusions et perspectives seront présentées Chapitre 6.



# Chapitre 2

## Bases bibliographiques et démarche adoptée

Ce chapitre fait un état de l'art des différentes méthodes d'analyse de la texture d'image rencontrées en télédétection et plus particulièrement dans le contexte des images THR et des applications forestières. La démarche adoptée pour la suite de ce travail est ensuite présentée.

### 2.1 Texture : définition

Si l'on reconnaît une texture au regard, il est cependant difficile de la définir. Ainsi il existe de nombreuses définitions dépendantes de l'appréciation et de la vision de chaque observateur. La texture peut être représentée par ce qui constitue une région d'un point de vue macroscopique, sa structure étant constituée par la répétition d'un motif suivant des règles spatiales (Tuceryan and Jain, 1998). Une région d'image a une texture constante si ses propriétés ou statistiques locales sont constantes ou varient faiblement (variations approximativement périodiques). La texture est une fonction de la variation spatiale des niveaux de gris. La texture peut être caractérisée par les trois ingrédients principaux que sont : la répétition d'un ordre local sur une large région (comparativement avec la "taille" de cette ordre); cet ordre consiste en un arrangement non chaotique d'éléments; ces éléments sont des entités uniformes ayant approximativement les mêmes dimensions sur l'ensemble de la région texturée (Tuceryan and Jain, 1998). De façon triviale, la nature et la dimension de ces entités dépendent directement de la résolution spatiale. L'information texturée sur une image de forêt est formée par l'alternance de pixels représentant la canopée éclairée, la canopée ombragée, les ombres projetées (au sol) et la réponse des autres éléments de la surface (sous-bois, sol...) (Ni

et al., 1999). Selon le rapport entre la résolution et la taille des arbres considérés, la texture peut aussi naître de la complexité interne de la canopée, alternances de feuilles, de branches et d'interstices. La géométrie de l'acquisition joue un rôle prépondérant dans la formation de la texture, notamment en terme d'orientation et d'amplitude des ombres. Les paramètres communs à tous les calculs de texture sont la taille de fenêtre utilisée ainsi que la bande spectrale à partir desquelles ils sont déterminés.

## 2.2 Texture THR et forêts

Pour la télédétection en milieu forestier, l'information texturale est principalement utilisée pour deux objectifs principaux correspondant à nos objectifs d'étude. La texture des images est utilisée pour estimer les principales variables de la structure forestière (Castillo et al., 2010; C.Song and Woodcock, 2003; Estes et al., 2010; Feng et al., 2010; Gomez et al., 2012; Guyon and Riou, 1996; Hyypä et al., 2000; Kayitakire et al., 2006; Ozdemir and Karnieli, 2011; Proisy et al., 2007; Song et al., 2010; Tuominen and Pekkarinen, 2005; Wolter et al., 2009; Wood et al., 2012; Wunderle et al., 2007) ou pour cartographier la structure forestière (Coburn and Roberts, 2004; Franklin et al., 2001; Murray et al., 2010; Regniers et al., 2013; St-Onge and Cavayas, 1997; Van Coillie et al., 2007; Vine et al., 1996) et ses changements (Boukir et al., 2013; Chehata et al., 2011, 2014; Huang et al., 2008).

De nombreuses études ont porté sur l'estimation des variables de structure forestière à partir d'images optiques depuis les débuts de l'imagerie satellitaire. Les méthodes consistent généralement à extraire une ou plusieurs variables prédictives à partir des images pour estimer des variables à l'échelle de l'arbre ou du peuplement qui sont utiles pour l'inventaire et la gestion des ressources forestières. Ces variables sont : la hauteur des arbres, le diamètre des couronnes, le diamètre des troncs, la densité, la surface terrière, le volume de la tige ou encore la biomasse (Feng et al., 2010; Kayitakire et al., 2006; Ozdemir and Karnieli, 2011; Proisy et al., 2007; Song et al., 2010). Ces études sont généralement basées sur un modèle de régression linéaire simple ou multiple. La précision dans l'estimation des variables du peuplement forestier dépend de la résolution spatiale de l'image. Hyypä et al. (2000) ont observé que les images satellites multispectrales de moyenne à haute résolution telles que SPOT-4 (résolution de 20m en multispec-

tral) ou Landsat TM (résolution de 30m en multispectral) conduisent à de faibles performances, en comparaison avec la photographie aérienne sub-métrique. En utilisant des images SPOT-5 (résolution de 5m en panchromatique et 10m en multispectral), Wunderle et al. (2007), Wolter et al. (2009), Castillo et al. (2010) ont pu estimer certains attributs des peuplements forestiers (tels que le diamètre de la couronne), en exploitant la texture d'image, avec une bonne précision.

Au cours de la dernière décennie, un nombre croissant d'images satellitaires multispectrales à Très Haute Résolution (THR), provenant de divers capteurs, sont disponibles comme Pléiades, Quickbird, Geoeye, WorldView ou Ikonos qui fournissent une résolution spatiale fine (0,5-1m en bande panchromatique, 2-4m en bandes multispectrales). Des études récentes (Barbier et al., 2010; Beguet et al., 2013, 2012, 2014; Couteron et al., 2005; Feng et al., 2010; Gomez et al., 2012; Kayitakire et al., 2006; Ozdemir and Karnieli, 2011; Proisy et al., 2007,?; Regniers et al., 2013; Ruiz et al., 2004; Song et al., 2010; Tuominen and Pekkarinen, 2005; Van Coillie et al., 2007) ont montré le potentiel de l'imagerie THR pour des applications d'inventaire forestier grâce à la forte relation existante entre la structure forestière et la texture de l'image, cela au niveau du peuplement. Dans ces travaux, de nombreuses représentations de la texture sont proposées. L'analyse de la texture à partir d'images satellites THR peut être divisée en trois grands types d'approche (Maillard, 2003) : les méthodes basées sur l'analyse fréquentielle (par exemple la transformée de Fourier ou par ondelettes), les méthodes basées sur l'utilisation de variogrammes, et les méthodes basées sur les matrices de co-occurrences (GLCM ou Grey Level Coocurrence Matrix).

Couteron et al. (2005), Proisy et al. (2007), Barbier et al. (2010) ont utilisé un périodogramme de Fourier 2D pour modéliser la texture de la canopée. Ruiz et al. (2004), Regniers et al. (2013), Van Coillie et al. (2007) ont utilisé des transformées en ondelettes pour l'analyse de la structure forestière. Les variogrammes sont aussi utiles pour explorer la relation entre la texture des images et la structure de la forêt (Franklin et al., 2001; Guyon and Riou, 1996; Song et al., 2010; St-Onge and Cavayas, 1997; Wulder et al., 1998). Leur principal inconvénient est l'exigence d'un bon modèle non-linéaire pour construire le variogramme avant d'extraire des indicateurs de texture comme le seuil, la portée et la pépite. Les GLCM (Section 2.3.3) fournissent des statistiques de second ordre définies par Haralick et al. (1973). Celles ci sont les attributs de texture les plus utilisés dans la littérature de télédétection forestière (Boukir et al., 2013; Chehata et al., 2011; Franklin et al., 2001). Elles fournissent de bonnes performances pour l'estimation

des paramètres de la forêt (Castillo et al., 2010; Kayitakire et al., 2006; Ozdemir and Karnieli, 2011; Wolter et al., 2009; Wunderle et al., 2007). De plus, leur implémentation reste moins complexe que d'autres représentations de la texture comme les transformées en ondelettes ou les filtres de Gabor et de nombreux logiciels de traitement d'images de télédétection les rendent aujourd'hui accessibles à tous. L'inconvénient majeur de cette représentation de la texture est le grand nombre de paramètres spatiaux (taille de la fenêtre de calcul, longueur du déplacement, orientation) ainsi que le nombre de niveau de quantification qui sont extrêmement difficiles à déterminer a priori étant donné la variabilité spatiale de la structure forestière au sein d'un peuplement et entre peuplements. Généralement ces paramètres sont fixés et par conséquent ils ne sont pas optimaux. Pour pallier cette limitation méthodologique, nous proposons une approche automatisée de sélection d'attributs de texture qui permet d'explorer une vaste gamme de paramétrages possibles. Cette approche, à notre connaissance, n'a pas été proposée jusqu'alors. Cette approche a un autre avantage : elle fournit un outil d'analyse de texture multi-échelle (en combinant des attributs de texture calculés avec des tailles de fenêtres différents) or il est bien connu que l'analyse de texture par des outils multi-échelle est plus optimale (Gonzales and Woods, 2008).

Enfin, l'impact des paramètres d'acquisition des images sur l'information texturale est une question cruciale. Les angles de visée du capteur et la position du soleil influencent fortement la texture des images en raison de l'interaction entre le rayonnement et la structure forestière. Ils déterminent notamment la longueur et l'orientation des ombres portées de la canopée sur le sol, la part de canopée éclairée ou ombragée et donc le rayon ou la longueur apparente des couronnes vues par le capteur. Barbier et al. (2010) a montré que les variations bidirectionnelles de la texture sont importantes et doivent être prises en compte en utilisant des images avec différentes configurations angulaires. De plus, la texture des images peut varier en fonction de la saison du fait de l'impact des changements phénologiques de la structure de la végétation (et notamment l'état de la végétation du sous bois). Ceci est donc à prendre en compte lorsque l'on mesure la texture à différents stades phénologiques (Culbert and al, 2009).

On comprend alors la nécessité d'adapter les descripteurs de la texture des images pour une date et un site donné, ceci renforçant l'idée qu'une optimisation des paramétrages des attributs de texture au cas par cas est primordiale. Après avoir passé en revue les principaux outils d'analyse de texture rencontrés dans

la littérature, nous détaillerons la démarche générale proposée dans ce manuscrit pour l'estimation des variables forestières et la cartographie des types de structure forestière à partir de la texture des images à très haute résolution.

## 2.3 Analyse de texture

Il existe de nombreuses approches pour caractériser la texture présente dans une image. L'étude s'est concentrée sur les attributs les plus couramment utilisés pour des thématiques forestières. Cependant, il semble important de s'ouvrir aux techniques employées dans d'autres domaines d'applications du traitement d'image.

### 2.3.1 Statistiques de texture du premier ordre

Les statistiques du premier ordre mesurent la vraisemblance d'observer une valeur de niveau de gris (luminosité du pixel) à un point donné au hasard. Elles dépendent uniquement des valeurs des pixels et ne tiennent donc pas compte des possibles relations de voisinage. Ces statistiques peuvent être calculées directement à partir de l'histogramme d'une image (Tuceryan and Jain, 1998) de façon globale (objet) ou locale (voisinage). Si la moyenne apporte une certaine information, c'est surtout la variance qui est employée notamment pour des résolutions relativement basses (supérieures à 5m) (Coburn and Roberts, 2004; Dionne et al., 1996; Guyon and Riom, 1996; Vine et al., 1996). Cette variance locale décrit la variabilité du signal au sein d'un voisinage donné. Deux textures homogènes différentes auront donc des variances différentes et pourront être distinguées avec ce critère. Le choix de la taille de la fenêtre de calcul est ici prépondérant.

Ces statistiques dépendent fortement de la résolution de l'image. Si la résolution est significativement plus fine que l'objet à observer alors les pixels voisins sont fortement corrélés et la variance locale est faible (sauf si on étend fortement le voisinage). Si la résolution est similaire à la taille des objets alors deux pixels voisins auront moins de chance d'avoir des comportements similaires et la variance augmente (Dionne et al., 1996; Vine et al., 1996). Enfin, si la résolution devient trop basse, le pixel est une intégration du signal sur différents objets, la variance locale diminue alors. D'autres statistiques du premier ordre sont aussi utilisées par Tuominen and Pekkarinen (2005) comme l'écart-type, le nombre de pixels uniques ou encore l'étendue des valeurs des pixels. La taille de la fenêtre

de calcul est l'unique paramètre de calcul des statistiques du premier ordre: elle peut varier selon le phénomène étudié par exemple elle est entre  $3 \times 3$  et  $8 \times 8$  pour Tuominen and Pekkarinen (2005) avec une résolution de 0.5m. Si les statistiques du premier ordre sont utiles pour décrire la variabilité locale, elles n'incluent pas de dépendance spatiale et ne permettent donc pas de caractériser de façon intrinsèque la structure de la texture.

### 2.3.2 Semivariogrammes

Les techniques d'analyse géostatistique constituent un intérêt grandissant pour l'analyse d'images et plus particulièrement pour l'analyse de texture (St-Onge and Cavayas, 1997). Plusieurs études ont montré l'utilité du semivariogramme pour quantifier la texture mais peu d'entre elles l'utilisent pour la cartographie. Le semivariogramme représente l'évolution de la variance en fonction de la distance entre deux points de la scène, ainsi il mesure la structure spatiale de la scène. Une étape de modélisation est nécessaire pour caler une courbe représentant au mieux le semivariogramme expérimental. En cas de stationnarité, les modèles les plus couramment utilisés sont exponentiels ou sphériques.

Cette courbe est alors caractérisée par son palier (sill) qui est de façon asymptotique la variance maximale observée dans la scène, sa portée (range) qui indique à partir de quelle distance la variance maximale est observée (ce qui représente la limite de la dépendance spatiale) et enfin l'effet de pépité (valeur non nulle de la limite du variogramme en zéro). Cet effet représente la variation entre deux mesures infiniment proches, et peut donc provenir de trois effets : une variabilité naturelle de la variable mesurée : il pourra par exemple prendre deux valeurs différentes si mesuré à deux instants différents ; une variabilité de l'instrument de mesure : la pépité mesure donc en partie l'erreur statistique de l'instrument de mesure ; un réel effet pépité : une variation spatiale brutale de la variable mesurée (le cas historique est le passage sans transition d'une pépité d'or à un sol ne contenant quasiment pas d'or). Cet éventuel effet n'est a priori pas exploité dans la littérature rencontrée.

Feng et al. (2010) utilisent le semivariogramme pour estimer le diamètre moyen des couronnes de différents peuplements lorsque la densité du peuplement est importante. Le semivariogramme est calculé le long d'un transect sur une ligne ou colonne de l'image (de façon aléatoire pour les cas de très forte densité et perpen-

diculairement aux rangs lorsqu'il y en a) ainsi que de façon matricielle avec les orientations N-S et E-O. Wulder et al. (1998) calcule les semivariogrammes sur des transects de 35 pixels. Si ces variogrammes peuvent être utilisés directement pour comparer des types de textures différents, l'information qu'ils apportent sur la magnitude de la variabilité et sur l'étendue de la dépendance spatiale est utilisée. pour déterminer la dimension optimale de la fenêtre de convolution ainsi que les limites définies par la résolution spatiale (Franklin et al., 1996). Enfin cette approche géostatistique est utilisée dans l'élaboration de modèles statistiques visant à simuler la complexité des couverts forestiers (C.Song and Woodcock, 2003; Song et al., 2010; Song and Woodcock, 2002, 2003).

### 2.3.3 Matrice de co-occurrence (Grey Level Cooccurrence Matrix ou GLCM)

Les statistiques de texture du second ordre sont les plus couramment rencontrées dans la littérature pour caractériser des textures naturelles (Coburn and Roberts, 2004; Franklin et al., 2001; Kayitakire et al., 2006; Murray et al., 2010; Wunderle et al., 2007). Elles peuvent être définies comme la vraisemblance d'observer une paire de valeurs (en niveaux de gris) aux deux extrémités d'un segment de taille, de position et d'orientation quelconque (Tuceryan and Jain, 1998). Elles sont les propriétés de paires de valeurs de pixels. Une illustration de la construction d'une GLCM est présentée figure 2.1.

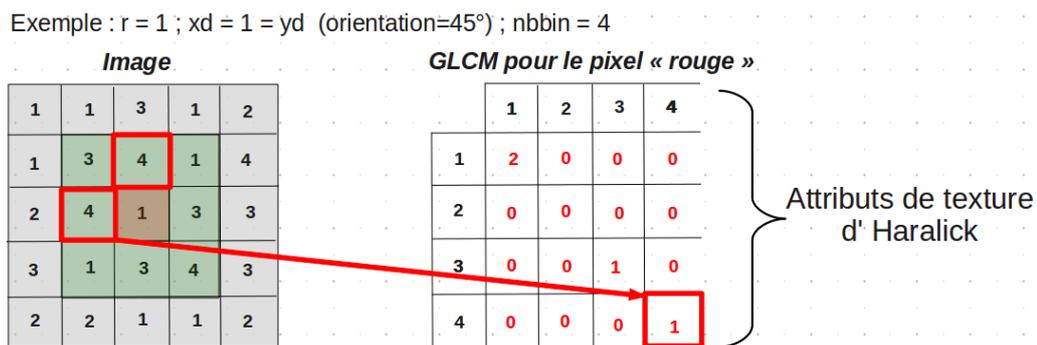


Figure 2.1 : Représentation schématique de la construction d'une GLCM. Les paramètres sont  $r$  pour le rayon de la fenêtre de calcul,  $xd$  et  $yd$  pour la longueur (et l'orientation) du déplacement,  $nbbin$  pour le nombre de niveaux de quantification. Est représenté en rouge le comptage d'une paire de niveau de quantification 4 dans la fenêtre de calcul (en vert) selon le déplacement et l'orientation définis dans cet exemple.

La matrice de cooccurrence permet d'estimer des propriétés statistiques de ce type : les nombreux descripteurs d'Haralick (au nombre de 14) qui en sont issus sont aujourd'hui les attributs de texture les plus utilisés en télédétection (Connors et al., 1984; Haralick et al., 1973). Cette matrice de co-occurrence est construite en comptabilisant le nombre de paires de pixels (pour toutes les combinaisons de paires de valeurs disponibles) séparés d'une même distance suivant une certaine orientation et ceci dans une fenêtre de taille à définir. Ce sont ces paramètres de distance et d'orientation qui introduisent des règles spatiales qui vont caractériser la texture. Chaque attribut dérivé de cette matrice est susceptible d'être représentatif d'une caractéristique précise de la texture étudiée. Les plus courants sont : l'énergie (energy), l'entropie (entropy), la corrélation (corr) et l'inertie (inertia). Dans la plupart des cas, ces attributs sont utilisés de façon conjointe. Le grand nombre de ces descripteurs peut entraîner une certaine redondance de l'information, ceci pouvant affecter les analyses. L'analyse en composantes principales appliquée aux multiples descripteurs pouvant être calculés à partir de la matrice de co-occurrence apparaît comme une bonne façon d'éviter cette redondance (Murray et al., 2010). Les attributs utilisés dans cette étude sont présentés dans le tableau 2.1.

La principale limite de cette méthode réside dans le choix de la taille de fenêtre. Cette dernière doit être suffisamment grande pour capturer l'ensemble de la structure de la texture tout en minimisant les effets de bords (Puissant et al., 2005). Pour calculer les descripteurs d'Haralick, Kayitakire et al. (2006) trouve que la taille de fenêtre optimale varie entre  $5*5$  et  $25*25$ ,  $15*15$  semblant la plus adaptée (ceci à partir d'images Ikonos à 1m). Le déplacement utilisé par Kayitakire et al. (2006) varie entre 1 et 3 (1 le plus souvent) selon les attributs. L'attribut de contraste est peu sensible au déplacement tandis que l'orientation affecte l'attribut de corrélation. L'orientation optimale dépend directement de la géométrie d'acquisition et donc de la formation des ombres, cette dernière peut changer en fonction du déplacement appliqué. Le contraste correspondant à l'alternance canopée éclairée canopée sombre, une fenêtre de  $5*5$  semble suffir. Ceci est confirmé par Wunderle et al. (2007) qui utilise aussi cette taille de fenêtre (après avoir testé des fenêtres de  $5*5$ ,  $11*11$ ,  $25*25$ ) avec des données Spot fusionnées à 2.5m. Coburn and Roberts (2004) combinent différentes tailles de fenêtre (entre  $5*5$  et  $15*15$  pour une résolution de 4m) pour réaliser une analyse multi-échelle. Cette approche lui permet d'obtenir de meilleurs résultats qu'en utilisant uniquement la taille de fenêtre supposée comme optimale. Ceci s'explique par le

fait qu'une zone ayant des variations spatiales faibles sera mieux caractérisée à l'aide d'une petite fenêtre tandis que pour les degrés de variation plus forts la fenêtre doit être plus grande. On comprend alors que la principale limitation des descripteurs du second ordre réside dans le choix a priori des paramètres de calcul et surtout de la taille de fenêtre. Une seule taille de fenêtre ne peut être adaptée à la caractérisation d'objets de tailles très différentes. L'approche multi-échelle semble pouvoir permettre de contourner ce problème afin d'appréhender différents types de textures.

Tableau 2.1 : Descriptions des attributs de texture issus des GLCM utilisés dans cette étude.

Attribut	Description	Formule
Energie (energy)	élevée pour une GLCM localement homogène	$\sum_{i,j} g(i,j)^2$
Entropie (entropy)	mesure du caractère aléatoire de la distribution des intensités	$-\sum_{i,j} g(i,j) \log_2 g(i,j)$
Correlation (corr)	Dépendance linéaire entre pixels voisins	$\sum_{i,j} \frac{(i-\mu)(j-\mu)g(i,j)}{\sigma^2}$
Inverse Difference (invdif)	mesure de l'homogénéité des valeurs de pixel, inversement proportionnel à l'inertie	$\sum_{i,j} \frac{1}{1+(i-j)^2} g(i,j)$
Inertie (inertia)	Valeurs fortes pour de fortes valeurs sur la diagonale de la GLCM	$\sum_{i,j} (i-j)^2 g(i,j)$
Cluster Shade (clushade)	moment d'ordre supérieur	$\sum_{i,j} ((i-\mu)+(j-\mu))^3 g(i,j)$
Cluster Prominence (cluspro)	moment d'ordre supérieur	$\sum_{i,j} ((i-\mu)+(j-\mu))^4 g(i,j)$
Haralick's Correlation (harcorr)	mesure de corrélation	$\frac{\sum_{i,j} (i,j)g(i,j) - \mu_i^2}{\sigma_i^2}$

Avec :

- $g(i,j)$  élément de la matrice de cooccurrence normalisée.
- $\mu = \sum_{i,j} i \cdot g(i,j) = \sum_{i,j} j \cdot g(i,j)$  la moyenne pondérée des valeurs des pixels.
- $\sigma = \sum_{i,j} (i-\mu)^2 \cdot g(i,j) = \sum_{i,j} (j-\mu)^2 \cdot g(i,j)$  (matrice symétrique) la variance pondérée des valeurs des pixels.

### 2.3.4 Approches fréquentielles

Les approches basées sur la transformée de Fourier (TF), la transformée en ondelettes ou les filtres de Gabor ont été peu utilisées pour caractériser la texture sur des scènes de forêt. La transformée de Fourier permet de passer d'une représentation de l'image dans le domaine spatial à sa représentation dans le domaine fréquentiel. Dans le cas de la forêt, les textures sont souvent quasi périodiques. Elles présentent une certaine régularité ou périodicité. Cette information est re-

dondante dans la représentation spatiale de l'image de texture. Par la nature du plan spectral, l'image de texture se retrouve réorganisée (et réduite) lorsqu'on adopte un tel plan pour l'observer. De ce fait, certaines informations sur la texture sont plus facilement accessibles dans le plan fréquentiel. Les analyses fréquentielles en 2D sont donc utiles pour décomposer une image par la fréquence et l'orientation de ces composantes. Couteron et al. (2005), Proisy et al. (2007), Barbier et al. (2010) ont montré l'utilité d'une telle analyse de texture notamment pour cartographier la biomasse de la mangrove. Ils utilisent la méthode FOTO (Fourier-based Textural Ordination) qui consiste à appliquer une analyse en composantes principales sur le spectre de Fourier (sur le module de la transformée de Fourier, la phase n'étant pas utilisée). Cette méthode leur permet d'estimer la biomasse même pour de fortes densités (là où d'autres méthodes saturent). Le seul paramètre de cette méthode est la taille de la fenêtre dans laquelle le spectre est calculé, pour une résolution de 1m (Ikonos), une taille de fenêtre entre 75m et 100m est préconisée dans Proisy et al. (2007). Si ces approches ont été peu utilisées jusqu'à présent, elles semblent adaptées à l'analyse d'images THR Couteron et al. (2005).

La TF reste un outil surtout utilisé pour son caractère global ou en première approximation. L'analyse locale par Fourier ne semble en effet pas suffisamment précise et insuffisamment riche en informations et en pertinence. Pour faire face aux difficultés d'analyse des textures naturelles, l'outil espace-fréquence et temps-échelle qui semble bien adapté est la transformée en ondelettes. Celle-ci permet de prélever des paramètres locaux dans le domaine spatial, dans le domaine fréquentiel, à des résolutions multiples ainsi que des paramètres inter-échelles. De plus, son mode d'investigation, du global vers le détail, possède une similitude avec le système visuel humain. L'utilité de la transformée en ondelettes et des filtres de Gabor (et autres filtres d'énergie) a été testé par Ruiz et al. (2004), Regniers et al. (2013) pour la segmentation et la classification d'images de forêts. D'autres méthodes multirésolutions permettant de mesurer des caractéristiques de l'image à des résolutions différentes existent: 1) la pyramide laplacienne et la pyramide gaussienne correspondant à des opérations de filtrage de taille fixe ou variable, 2) les transformées de Gabor et Wigner-Ville utilisées pour discriminer ou segmenter les textures (LeBris, 2012), 3) les fractales pour déterminer la dimension fractale à partir d'outils multi-échelles.

## 2.3.5 Autres approches

### 2.3.5.1 Basées sur la détection de primitives

Certaines analyses de texture se basent sur la détection de primitives et de leurs arrangements dans l'image. C'est le cas de la méthode TPN (Triangular Primitive neighborhood) proposée par Hay et al. (1996). Cette méthode hiérarchique permet d'identifier la position de chaque arbre en définissant un pic représentant le centre de sa canopée. Une triangulation de Delaunay est ensuite appliquée afin d'obtenir des régions triangulaires, ces dernières représentent visuellement les différences d'arrangement et de densité des différents groupes de trois arbres, elles sont alors considérées comme les primitives de la texture de l'image. Trois indicateurs principaux sont extraits : la surface du triangle, la moyenne et la variance des pixels contenus dans ce triangle. D'autres méthodes utilisent les primitives afin de réaliser par exemple des tessellations de Voronoi (Tuceryan and Jain, 1998). Il est à noter que ce type de traitements est difficile à mettre en œuvre, de plus ils sont coûteux en temps de calcul.

### 2.3.5.2 Basées sur la morphologie mathématique

La morphologie mathématique est aussi rencontrée dans les articles de télédétection, Huang et al. (2008) ont souligné son intérêt pour la discrimination d'espèces dans un écosystème de mangrove à partir d'images THR Ikonos. La morphologie mathématique est basée sur deux opérations élémentaires que sont la dilatation et l'érosion. En combinant ces dernières sont obtenues les opérations d'ouverture et de fermeture. Huang et al. (2008) obtient des résultats intéressants avec la dualité des opérateurs OFC (ouverture par reconstruction suivie de fermeture par reconstruction) et CFO (fermeture par reconstruction suivie d'une ouverture par reconstruction). Ceux-ci lui permettent d'obtenir de meilleurs résultats de classification qu'avec les attributs GLCM. Ces attributs sont d'autant plus efficaces s'ils sont couplés avec des attributs spectraux. Cependant, cette méthode n'ayant pas été testée pour d'autres études, il est difficile de prévoir son utilité pour la discrimination de différents types de texture.

## 2.4 Démarche adoptée

Pour répondre aux deux objectifs principaux de ce travail de thèse, à savoir l'estimation des variables forestières et la cartographie des types de structure forestière à partir de la texture des images à très haute résolution spatiale, nous avons choisi d'utiliser les attributs de textures issus des matrices de co-occurrence, outil d'analyse de texture de référence en télédétection forestière. La principale limite de cet outil d'analyse de texture étant le grand nombre de paramètres spatiaux à déterminer, le grand défi de ce travail est de palier ce problème. Pour ce faire, aussi bien pour l'objectif d'estimation des variables forestières que pour la cartographie des types de structure forestière, la méthodologie développée consiste à calculer en amont, sur un nombre réduit d'échantillons d'images, un grand nombre d'attributs de texture en balayant une vaste gamme de paramétrages possibles. La sélection automatique des meilleurs attributs et combinaisons d'attributs pour une question donnée permet d'une part d'optimiser les paramétrages de ces attributs et d'autre part d'exploiter au maximum le pouvoir explicatif de la donnée en combinant ces attributs. Or ces deux types de données sont souvent acquises simultanément avec le même satellite et elles diffèrent en résolutions spatiales (généralement sub-métrique en panchromatique et 4 fois plus faible en multispectral) et spectrale (une large bande en panchromatique et plusieurs bandes étroites et différentes en longueurs d'ondes en multispectral). Les études existantes ne prennent généralement en compte qu'une seule résolution spatiale, travaillant soit sur la donnée panchromatique soit sur la donnée multispectrale. A notre connaissance, seulement Wolter et al. (2009) a combiné des attributs de texture issus des deux résolutions, mais en utilisant des variogrammes et non pas des matrices de co-occurrence. Nous allons donc dans ce manuscrit explorer le potentiel du croisement des informations issus des deux résolutions. L'utilisation jointe de la texture issue des deux résolutions (panchromatique et multispectrale) couplée à la combinaison d'attributs de texture calculés avec des tailles de fenêtres différentes permet, de manière automatique et implicite, d'obtenir des descripteurs multi-échelle. Une attention particulière est portée sur le caractère opérationnel des méthodologies proposées, l'objectif étant, à terme, de fournir aux gestionnaires forestiers des outils d'aide à la décision pour la gestion forestière.

Afin d'explorer le potentiel des images THR pour la caractérisation de la structure forestière, des régressions linéaires multiples entre chaque variable de structure forestière et un ensemble d'attributs de texture optimal sélectionné sont appliquées. Il s'agit de montrer les relations entre la texture des images et les

variables de la structure forestière et aussi de d'établir le meilleur modèle de prédiction de chacune de ces variables forestières. Différentes méthodes de sélection de variables pour la régression linéaire multiple ont été testées, une attention particulière a été portée au problème de la colinéarité entre les variables (c'est à dire les attributs image). Une nouvelle approche de sélection de variables pour la régression linéaire multiple a été développée afin d'obtenir des modèles les plus performant possibles en terme de prédiction en s'assurant d'éviter le problème de la colinéarité entre les variables. Cette partie est détaillée dans le chapitre 4 de ce manuscrit.

Pour la cartographie des types de structure forestière fondée sur une classification d'image, une stratégie à visée opérationnelle est développée. Nous proposons une approche hiérarchique qui consiste dans un premier temps à discriminer les différents stades de mutation de la forêt (coupes rases, reboisements en cours ou récents) des stades forestiers à proprement parler (qui correspondent aux peuplements dont l'âge est supérieur à environ trois ans c'est à dire un âge où les arbres sont plus grands que la végétation du sous bois et par conséquent visibles dans l'image). Une fois cette première stratification réalisée, la seconde étape de cette approche hiérarchique se concentre sur la discrimination des grandes étapes d'évolution des peuplements forestiers. Cette approche est basée comme précédemment sur une sélection en amont d'un ensemble d'attributs de texture issus des deux résolutions. Les forêts aléatoires (Random Forest (RF)), classifieur d'ensemble, sont utilisés pour la classification et la sélection de variables. La sélection des attributs de texture est ici réalisée grâce à l'importance de variables obtenue par les forêts aléatoires, les  $n$  meilleurs attributs sont ensuite calculés sur l'image entière avant d'être passés en entrée du classifieur. Les forêts aléatoires sont ensuite utilisées pour réaliser la classification des images et produire une cartographie fine de la structure forestière. Un des intérêts principaux de ce classifieur est le concept de marge d'ensemble qui mesure la difficulté à classer une instance. Ces cartes de marge permettent une innovation thématique importante, elles donnent à l'opérateur humain une information sur la fiabilité de la classification et d'autre part elle permettent de détecter des erreurs d'étiquetage (ces dernières peuvent être fréquentes dans un contexte complexe comme le milieu forestier). Cette partie est détaillée dans le chapitre 5 de ce manuscrit.

Afin d'évaluer les méthodologies proposées et le potentiel des images THR pour caractériser la structure forestière, deux sites d'études et 4 jeux d'images

(panchromatique et multispectral) acquis à différentes saisons seront utilisés dans cette étude. Ceci permettra d'évaluer la robustesse des méthodologies développées, l'apport thématique des images THR mais aussi la quantification des erreurs aussi bien pour l'estimation des variables de structure forestière que pour la cartographie des types de structure forestière.



# Chapitre 3

## Données

### 3.1 Sites d'étude

Les deux sites d'étude sont situés dans le sud-ouest de la France, dans la plus grande forêt Européenne de pin maritime (*Pinus pinaster* Ait.), la forêt des Landes de Gascogne (Figure 3.1). Cette forêt couvre environ un million d'hectares dans une zone au relief plat à l'exception des dunes côtières.

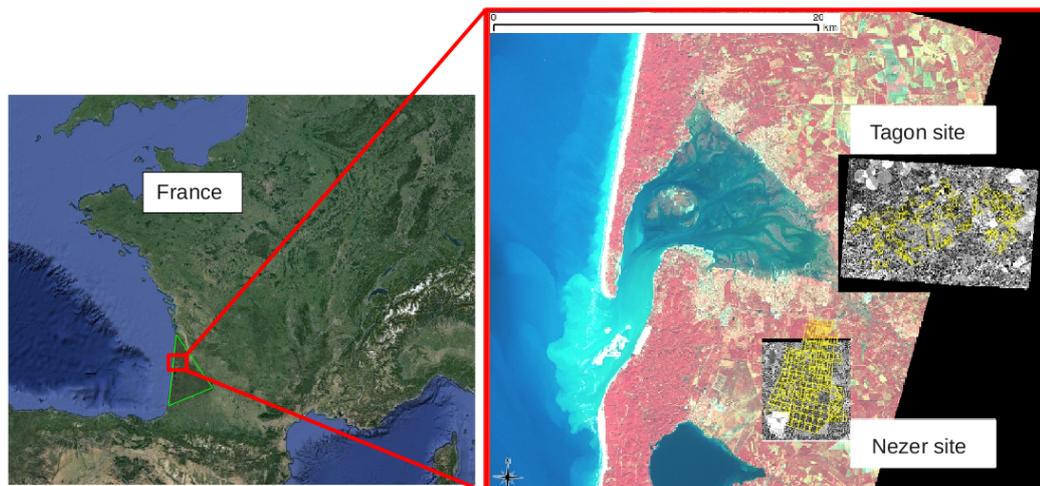


Figure 3.1 : Localisation des sites d'étude. Forêt des Landes de Gascogne (triangle vert), Aquitaine, France.

Cette forêt se compose de peuplements de pin maritime équiens qui sont gérés de manière intensive. Les pins sont issus de semis ou de plantations, en rangées généralement séparées de 4m. Les techniques de régénération ont changé ces dernières années. A l'heure actuelle, les arbres sont généralement plantés

régulièrement le long des lignes avec une densité assez faible :  $\sim 1000$  arbres/ha, soit un espacement fixe sur les rangs de l'ordre de 2,5 m. Auparavant, les arbres étaient toujours semés avec une densité supérieure sur les lignes :  $\sim 5000$  arbres/ha, c'est à dire un espacement moyen dans la rangée de l'ordre de 0,5 m. Les arbres sont périodiquement éclaircis après débroussaillage de la végétation de sous-bois. La coupe rase se produit surtout dès que les pins ont atteint l'âge de 50 ans.

Le site de Nezer (Figure 3.2), dont la superficie est d'environ 60km<sup>2</sup>, a une gestion forestière relativement uniforme, car il ne couvre que deux grandes exploitations sylvicoles. La taille des peuplements est grande, avec une superficie moyenne d'environ 12 ha pour un maximum de 50 ha. Les peuplements sont pour la plupart rectangulaires et souvent entourés par des pare-feu, des pistes. Les rangées d'arbres sont orientées Ouest-Est dans la plupart des peuplements.

Le site du Tagon (Figure 3.3), dont la superficie est d'environ 80 km<sup>2</sup>, est plus complexe et plus hétérogène que celui de Nezer, notamment car ses peuplements sont gérés par de nombreux propriétaires forestiers. Les peuplements sont plus petits (surface moyenne  $\sim 7$  ha, au maximum  $\sim 40$  ha) et présentent une grande variabilité en termes de géométrie, d'orientation des rangs, et de structure forestière.

Les deux sites sont très représentatifs de toute la diversité du massif forestier des Landes de Gascogne en termes de structures forestières, de composition des espèces de la végétation de sous bois mais aussi en termes de sylviculture et de gestion forestière. Le site de Nezer ayant été très touché par la tempête Klaus de 2009, il a été nécessaire de changer de site expérimental car aujourd'hui, il n'y a plus toute la diversité des structures forestières, les peuplements les plus âgés ( $> 20$ ans) ayant été très endommagés et remplacés.

## 3.2 Données images

Des images satellitaires provenant de deux capteurs à très haute résolution spatiale (Quickbird et Pléiades) ont été utilisées. Ces images ont des résolutions spatiales similaires pour une bande spectrale donnée :  $\sim 0.5-0.6$  m pour le panchromatique (Pan) et  $\sim 2-2.5$ m pour le mode multispectral (MS).



Figure 3.2 : Site de Nezer, localisation des peuplements et des placettes de mesure ( $n=12$ , points verts). Image de fond Quickbird panchromatique.

Pour le site de Nezer, une image Quickbird acquise le 6 octobre 2003 a été traitée et analysée. Ceci a permis de préparer l'arrivée des premières images Pléiades. Pour le site du Tagon, trois images ont été acquises respectivement le 26 juin 2012, le 8 août 2012 et le 20 février 2013.

Toutes les images ont été livrées géo-référencées et ortho-rectifiées avec une très grande qualité (avec une précision en x y inférieure à 0.5m (RMSE)), leurs caractéristiques sont présentées Table 3.1. La résolution radiométrique des images Pléiades est de 12bits, une image (celle du mois d'Août) ayant été livrée par erreur avec un codage en 8bits, nous avons passé toutes les images en 8bits dans un souci d'uniformisation. Cette précision radiométrique s'est avérée être suffisante pour nos applications où les indices de texture sont calculés après recodage

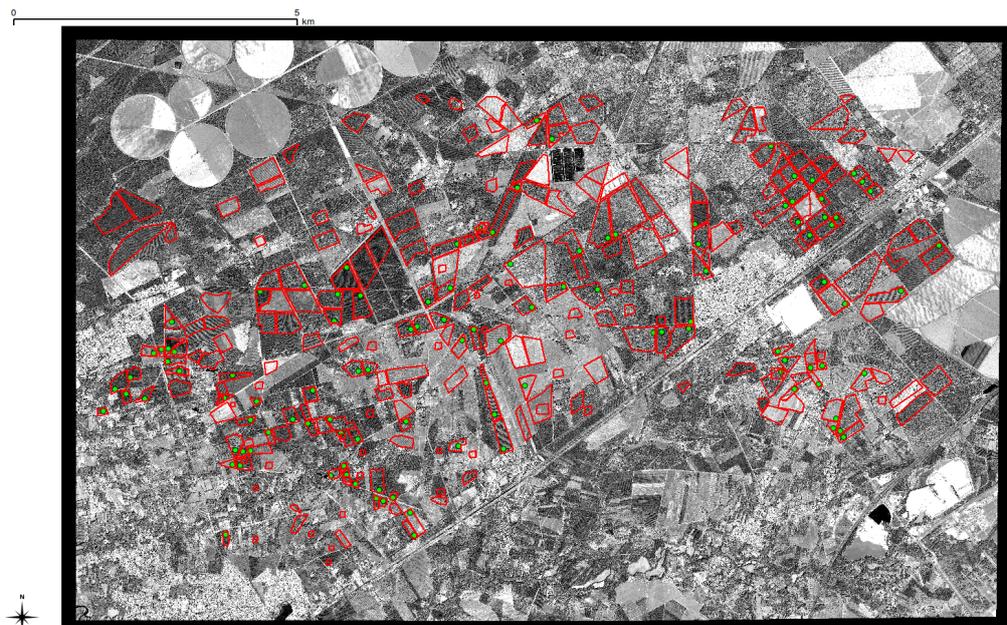


Figure 3.3 : Site de Tagon, localisation des peuplements de pin étudiés et des placettes de mesure ( $n=111$ , points verts). Image de fond Pléiades pancromatique du mois de juin.

sur 8 classes radiométriques.

Cet ensemble d'images fournit une répétition des observations satellitaires dans différentes conditions en termes de :

- stade phénologique de la végétation (été / automne / hiver);
- élévation solaire (haute / basse);
- angles de vue (oblique / très oblique).

Pour l'image Quickbird sur le site de Nezer acquise en octobre, la végétation du sous-bois était en fin de sénescence avec donc peu de feuilles vertes quelque soit sa composition en espèces. La contribution de la réflectance du sous bois et de sa variabilité sur l'image et sa texture sont donc probablement très faibles. Au contraire, pour le site du Tagon et les deux images d'été, le taux de recouvrement des feuilles vertes de la végétation du sous bois est à son maximum. L'influence de cette végétation sur la réflectance peu varier fortement dans l'espace en fonction des espèces présentes (fougères, ajoncs, molinie et autres) et de leurs propriétés spectrales. On attend donc un effet non négligeable de la végétation du sous bois sur la réflectance mesurée par le capteur et par conséquent sur la texture des

Tableau 3.1 : Paramètres d'acquisition des images.

	QuickBird 06/10/2003	Pléiades 26/06/2012	Pléiades 08/08/2012	Pléiades 20/02/2013
PAN-MS résolution (m)	0.62 - 2.48	0.5 - 2	0.5 - 2	0.5 - 2
Codage (bits)	11	8	8	8
Panchromatique	450-900 nm	470-830 nm	470-830 nm	470-830 nm
Bleu (B)	450-520 nm	430-550 nm	430-550 nm	430-550 nm
Vert (V)	520-600 nm	500-620 nm	500-620 nm	500-620 nm
Rouge (R)	630-690 nm	590-710 nm	590-710 nm	590-710 nm
Proche Infra Rouge (PIR)	760-900 nm	740-940 nm	740-940 nm	740-940 nm
Angle solaire azimuthal	157.3	147.3	154.3	159.1
Angle d'élévation solaire	33.1	66.2	59.3	32.4
Angle de vue azimuthal	42.5	179.9	180.0	180.0
Angle de vue zenithal	20.0	24.8	13.5	19.9
Site d'étude	Nezer	Tagon	Tagon	Tagon

images. Enfin, l'image d'hiver sur le site du Tagon permettra de comparer les performances des images Pléiades entre les deux saisons. L'élévation solaire étant plus basse en hiver, les ombres portées sont plus longues et plus importantes et devraient donc avoir un effet très différent sur la formation de la texture, par contre, comme évoqué précédemment, la végétation du sous bois à cette saison a une surface foliaire verte minimale, avec les espèces herbacées complètement sèches, par conséquent elle a un effet minime sur la réflectance du couvert mesurée par satellite. Ces considérations sur la phénologie du sous bois en forêt de pin maritime sont fondées sur les travaux de Yauschew-Raguenes et al. (2008) et Yauschew-Raguenes (2012).

### 3.3 Informations forestières de référence

#### 3.3.1 Mesures dendrométriques in situ sur les deux sites

Pour les deux sites d'étude, quatre variables de structure forestière ont été mesurées in situ : le diamètre des troncs à hauteur de poitrine ( $Dbh$ ) c'est à dire à 1,30m, la hauteur des arbres ( $Ht$ ), le diamètre des couronnes ( $Cd$ ) et la densité du peuplement ( $Nah$ ). L'espacement entre les arbres ( $Sp$ ) a été calculé à partir de la densité du peuplement en utilisant une fonction non linéaire (cf. Equation 3.1), cette dernière se base sur l'hypothèse que chaque arbre est situé sur le sommet d'un triangle équilatéral.

Les méthodes et techniques de mesure in situ des variables forestières sont détaillées en annexe A.

$$Sp(m) = \sqrt{\frac{20000}{\sqrt{3} * Nah(arbres/ha)}} \quad (3.1)$$

Pour le site de Nezer, les mesures ont été réalisées sur 12 peuplements en mars 2004, c'est à dire à la fin de l'année de croissance 2003. Les arbres échantillonnés dans chaque peuplement se trouvaient dans une placette carrée de 80m x 80m (6400m<sup>2</sup>) représentative de la structure forestière d'une région englobante de 200m x 200m. En chaque placette, la hauteur du sommet, le Dbh, le diamètre des couronnes sont mesurés respectivement sur 30, 30 et 10 pins échantillonnés aléatoirement. Les techniques de mesure sont identiques à celles employées sur le Tagon (cf Annexe A). Le protocole est détaillé dans Guyon et al. (2005) et évoqué dans Garestier et al. (2009).



Figure 3.4 : Carte des âges des peuplements sur le site de Nezer en 2003.

L'âge exact des peuplements fourni par les gestionnaires forestiers est connu

sur l'ensemble du site, la carte des âges est présentée dans la figure 3.4. Toutes les données ont été fournies par D. Guyon, responsable scientifique du dispositif expérimental installé par l'INRA à Nezer pour ses travaux en télédétection (cf par exemple Garestier et al. (2009), qui utilise le même jeu de données).

Pour le site du Tagon, la campagne dendrométrique a été faite spécifiquement pour mon étude. Les mesures ont été réalisées sur 111 peuplements durant l'été 2012, à la fin de la période de croissance des pins. Nous avons conçu un protocole spécifique pour cette campagne de mesure, il est détaillé en Annexe A.

Les arbres échantillonnés dans chaque parcelle sont inclus dans une placette de 400m<sup>2</sup> représentative de la structure forestière d'une région englobante de 80m x 80m. En chaque placette, les  $n$  arbres les plus proches du centre sont mesurés :  $n=10$  pour  $Dbh$ ,  $n=5$  pour  $Ht$  et  $Cd$ . L'âge exact n'est pas connu, il est estimé par dénombrement des niveaux de branches, très imprécis quand les branches les plus âgées ont disparus. Les mesures de terrain proprement dites ont été faites par l'unité expérimentale Forêt-Pierroton de l'INRA Bordeaux (B. Issenhut).

La campagne de mesure terrain effectuée sur le site du Tagon a été guidée par un travail en amont réalisé par SIG (Système d'Informations Géographique) à l'aide de photographies aériennes et d'une image Geoeye acquises en 2009 (cf. Section 3.3.2). L'objectif était ici d'échantillonner uniformément toute la gamme de variation des variables de la structure forestière présente sur le site. Les mesures ont été faites sur les peuplements âgés de 4 ans et plus, en effet, pour les peuplements plus jeunes, la fraction de couvert forestier est très faible et la végétation de sous bois est souvent plus haute que les pins.

Les gammes de variation des variables de la structure forestière pour les deux sites sont présentées dans le tableau 3.2. La figure 3.5 permet d'observer les relations allométriques entre les variables  $Cd$ ,  $Ht$  et  $Sp$ , soit leurs corrélations intrinsèques à l'espèce, liées à la croissance des arbres, la fertilité du site et opérations sylvicoles effectuées. On peut observer que les relations entre les différentes variables forestières sont similaires pour les deux sites.

La campagne de mesures sur le site du Tagon a permis d'échantillonner les peuplements des plus jeunes aux plus vieux et ainsi piéger toute l'étendue possible des variables forestières pour le pin maritime tandis que les données pour le site

de Nezer ne permettaient pas de prendre en compte les plus jeunes peuplements notamment ceux dont l'âge est inférieur à 13 ans.

Tableau 3.2 : Gamme de variation des variables forestières étudiées (Nezer : 12 peuplements, Tagon : 111 peuplements). Pour Cd, Ht et Dbh, les valeurs moyennes sont présentées.

		Cd(m)	Ht(m)	Nah(arbres/ha)	Sp(m)	Dbh(m)	Age(an)
<b>Nezer</b>	Min	2.95	9.3	189	3.04	0.15	13
	Max	7.81	24.7	1253	7.81	0.46	51
<b>Tagon</b>	Min	0.77	1.7	150	1.31	0.02	4
	Max	10.69	26.2	6729	8.77	0.56	68

### 3.3.2 Photo-interprétation sur le site du Tagon

#### 3.3.2.1 Photo-interprétation pour la campagne terrain

Dans un premier temps, afin de préparer au mieux la campagne de mesures in situ sur le site du Tagon, une étape de photo-interprétation visuelle a été réalisée. Celle-ci a porté sur les images satellites THR ou aériennes les plus récentes disponibles (avant l'acquisition des images Pléiades) : image Geoeye (16 février 2009, 40cm de résolution), orthophotographie IGN (couleur, 50cm, juillet 2009), orthophotographie INRA (IRC, 15cm, 26 février 2009). L'objectif de ce travail était d'échantillonner de manière uniforme les variables de la structure forestière sur l'ensemble du site. Pour cela, six classes de la structure forestière ont été créées, représentatives de la taille des couronnes des arbres, leurs caractéristiques sont présentées dans le tableau 3.3. Ces classes conduisent à une stratification de la structure forestière, ce qui a permis de sélectionner à l'avance un ensemble de parcelles en prenant le même nombre dans chacune des classes pour ensuite faire les mesures in-situ sur ces parcelles.

Les classes ayant été localisées sur les images de 2009, un ajustement à la situation 2012 a été fait de manière à prendre en compte les coupes rases et reboisements récents réalisés entre temps (visite terrain) et donc d'éventuels passages de la classe C1 à C2. Cette étape a permis d'assurer un échantillonnage uniforme sur l'ensemble du site, l'opérateur sur le terrain étant guidé par la localisation des parcelles d'intérêt via SIG. Un des objectifs de cette campagne de mesure était d'appréhender la variabilité intra-parcellaire en faisant des mesures dans 3 placettes pour chaque parcelles d'intérêt, malheureusement, faute de moyens,

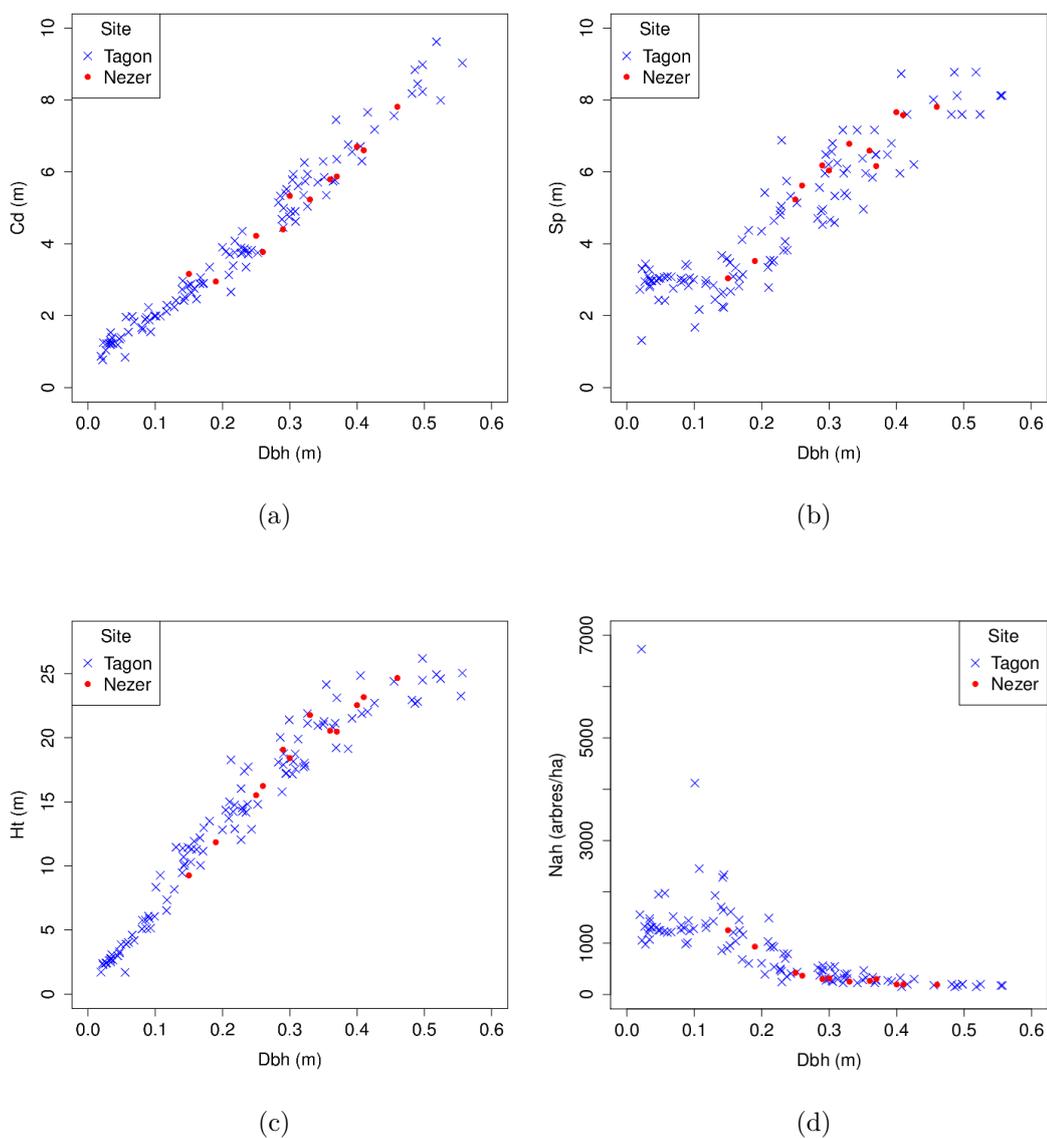


Figure 3.5 : Variables forestières mesurées sur les deux sites : diamètre moyen des couronnes ( $Cd$ ), hauteur moyenne des arbres ( $Ht$ ), espacement entre les arbres ( $Sp$ ) et densité des tiges ( $Nah$ ) en fonction du diamètre des troncs moyen ( $Dbh$ ).

nous n'avons pas réalisé cet objectif. Une seule placette de mesure par parcelle a été considérée, cet objectif est néanmoins d'un grand intérêt pour approfondir cette étude et l'idée est à conserver pour des campagnes de mesures à venir.

*Tableau 3.3 : Descriptif des classes de structures forestières.*

Classe 1	Coupes rases ou reboisements récents, âge du peuplement $\leq 3$ ans (saison 2012 comprise)
Classe 2	Rangs / arbres et ombres discernables couvert non fermé sur l'inter rang / $Cd < 1$ m
Classe 3	Rangs serrés / arbres confondus $Cd \sim 1.5-2.5$ m ( $< 3$ m)
Classe 4	Rangs moins visibles / arbres discernables densité généralement forte / $Cd \sim 3-4$ m ( $< 4.5$ m)
Classe 5	Rangs rarement visibles / arbres discernables densité généralement moyenne / $Cd \sim 4.5-6$ m ( $< 6.5$ m)
Classe 6	Rangs rarement visibles / arbres discernables densité généralement faible / $Cd \sim 6.5$ m et plus

Cette préparation de la campagne nous a permis d'avoir une base de données de grande qualité avec un grand nombre de parcelles ( $n=111$ ). De plus, la campagne terrain a été réalisée durant l'été 2012, soit au même moment que l'acquisition des premières images Pléiades, ceci permettant d'avoir un jeu de données complet: images et données terrain.

La figure 3.6 représente les relations entre variables forestières avec la classe d'appartenance donnée a priori via la photo interprétation. Les mesures in situ ont été réalisées pour les classes de 2 à 6 (qui correspondent aux peuplements dont l'âge est supérieur à 3 ans). Ces figures permettent d'observer que l'échantillonnage est bien régulier pour les différentes classes. On observe aussi un certain recouvrement entre les classes, ceci rappelant la nature continue de la structure forestière, les limites des classes sont floues. Cela illustre aussi la finesse de la photo-interprétation visuelle, puisque seulement deux peuplements ont été mal classés (classe 4 au lieu de classe 3); cette erreur d'affectation a été corrigée par la suite.

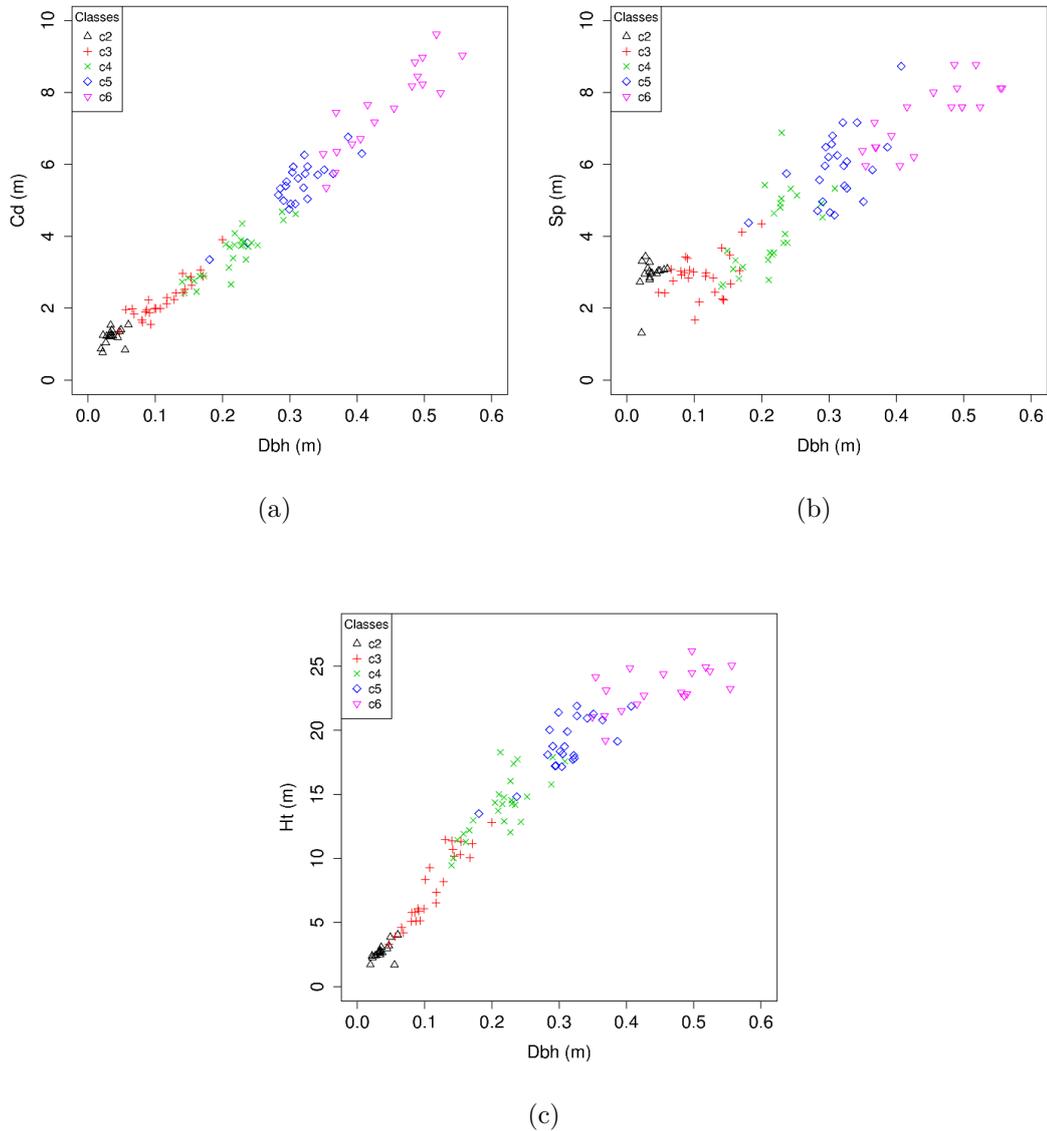


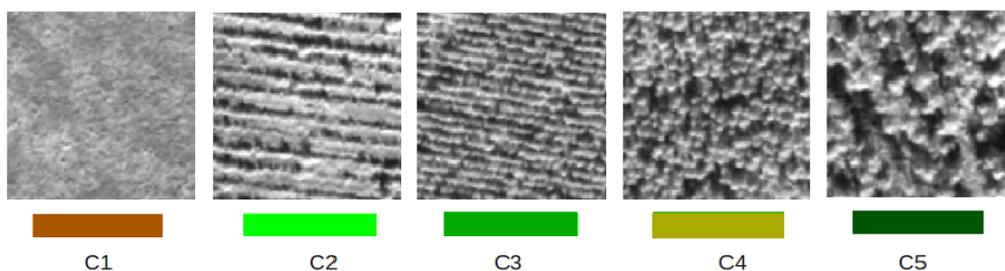
Figure 3.6 : Variables forestières mesurées et classes d'appartenance : diamètre des couronnes ( $Cd$ ), hauteur des arbres ( $Ht$ ) et espacement entre les arbres ( $Sp$ ) en fonction du diamètre des troncs ( $Dbh$ ).  $C2$  à  $C6$  : classes de structure a priori (cf tableau 3.3) c'est à dire données par photo-interprétation avant la campagne de terrain. Noter l'erreur d'affectation sur 2 peuplements ( $C4$  au lieu de  $C3$ ) corrigée par la suite.

### 3.3.2.2 Photo-interprétation pour la classification des structures forestières

Sur cette base, en vue du travail de classification d'image (cartographie des types de structure forestière), un ensemble de 200 peuplements supplémentaires ont été étiquetées via photo interprétation des images Pléiades panchromatiques. Ceci afin de constituer une base de données de référence sur le site du Tagon. Etant

donnée la complexité de la structure forestière, cette étape de photo-interprétation est laborieuse et délicate, en effet, si les échantillons présentés dans la figure 3.7 sont bien au centre des classes, les frontières entre ces classes sont floues et il est difficile pour un certain nombre de parcelles (notamment pour les peuplements plus âgés) de donner une étiquette sans une certaine incertitude.

Après les premiers tests pour la cartographie des types de structure forestière, étant donnée la grande confusion entre les deux dernières classes 5 et 6, la stratification de la structure forestière a été simplifiée, les classes 5 et 6 présentées 3.3 ont été regroupées. Les 5 classes de structure forestière qui seront considérées dans la suite de ce travail sont présentées dans la figure 3.7.



*Figure 3.7 : Imagerie panchromatique et classes de la structure forestière associée.*  
*C1 = pas d'arbres visibles (coupes rases, reboisement en cours ou récent), âge des peuplements  $\leq 3$ ans;*  
*C2 = arbres visibles, rangs visibles,  $Cd < 1m$ ; C3 =  $Cd \in [1m;2.5m]$ , rangs visibles;*  
*C4 =  $Cd \in [2.5m;4.5m]$ , rangs rarement visibles, densité plus faible;*  
*C5 =  $Cd \geq 4.5m$ , rangs non visibles, densité faible.*

# Chapitre 4

## Estimation des variables forestières à partir de la texture THR par modélisation statistique

*Ce chapitre se focalise sur l'estimation des variables de la structure forestière par modélisation statistique. Un état de l'art des méthodes de sélection par régressions linéaires est détaillé. La méthodologie proposée est ensuite exposée. Les résultats sont présentés sur les deux sites d'études en utilisant différents jeux de données THR disponibles. Des communications et publications ont fait suite à ce travail (Bequet et al., 2012, 2014,?).*

### 4.1 Introduction

L'objectif général de ce chapitre est de proposer et évaluer une méthode d'estimation des variables forestières utiles à inventaire de ressource forestière en exploitant au mieux le potentiel de la texture d'images THR (aux deux résolutions panchromatique et multispectral). Ces variables sont : la hauteur des arbres, le diamètre des couronnes, la densité, le diamètre des troncs et l'espacement entre les arbres. Pour cela, nous voulons explorer tout le potentiel de la texture des images THR : la bande panchromatique (avec une résolution submétrique) ainsi que les bandes multisp spectrales (avec une résolution spatiale moins importante). Les matrices de co-occurrence sont utilisées pour analyser la texture. Nous avons choisi d'estimer les variables forestières à l'aide de modélisation statistique par régression linéaire simple et multiple. L'évaluation est faite en forêt de pin maritime du massif landais et sur les images Quicbird et Pléiades, néanmoins, la méthode développée se veut générique.

Afin de palier au problème du paramétrage spatial des attributs de texture issus des matrices de co-occurrence, l'idée est de ne pas faire d'hypothèse a priori sur le paramétrage spatial et de calculer un très grand nombre d'attributs de texture (les 8 attributs les plus rencontrés en forêt) en balayant une vaste gamme de paramétrages possibles. Pour la régression linéaire multiple, nous devons développer des modèles parcimonieux (avec un nombre limité d'attributs image comme variables prédictives) tout en réduisant au minimum la colinéarité au sein de ces modèles, cela de façon automatique. La méthodologie proposée se devra d'être robuste aux conditions d'observation satellitaire qui influent sur la réflectance et la texture des images.

Les développements méthodologiques proposés sont donc l'automatisation de la sélection du modèle le plus performant, ainsi que l'automatisation de la sélection de combinaisons optimales des résolutions spatiales et bandes spectrales disponibles. Pour la méthode de sélection de variables en cas de régression multiple, nous proposons une nouvelle méthode, nommée Random PRiF adaptée au problème de la colinéarité. Les performances de cette nouvelle méthode seront comparées à des méthodes classiques.

Après avoir présenté un état de l'art sur les méthodes de sélection de variables pour la modélisation linéaire multiple, nous détaillerons la méthodologie proposée. Une partie présentera les résultats principaux obtenus lors de cette étude avant de conclure.

## 4.2 Méthodes de sélection de variables par régression linéaire

La plupart des articles rencontrés combine diverses informations extraites des images, qu'elles soient spectrales ou texturales. L'utilisation de régressions linéaires multiples permet de mettre à profit la possible complémentarité des descripteurs image pour la prédiction de variables de structure forestière. Cependant, le nombre d'attributs considéré peut être très important et affecter aussi bien la qualité que la compréhension des résultats. Une sélection des attributs les plus pertinents est donc essentielle. Différentes méthodes de sélection ont ainsi été testées par Haapanen and Tuominen (2008) telles que les algorithmes génétiques, des

sélections itératives ascendantes (Forward selection) et une version dérivée de l'algorithme du simplexe. Il est important de noter que la majorité des études ne cherche pas à sélectionner les attributs a priori mais plutôt à tester différentes combinaisons afin de trouver l'optimale tel le cas des régressions multiples faites pas à pas. Les analyses en composantes principales sont aussi utilisées pour synthétiser l'information, l'inconvénient étant la difficulté d'interpréter l'information contenue dans les premières composantes.

Au regard des différentes études menées, la combinaison d'informations apparaîtrait comme essentielle. Il reste à choisir entre sélectionner les attributs a priori et ainsi alléger les traitements ou bien tester toutes les combinaisons possibles de façon exhaustive. Si cette dernière approche semble être la plus adaptée à un traitement automatisé, elle pose cependant le problème de la détection et de la minimisation de la colinéarité entre les différentes variables image.

### 4.2.1 Détection de la colinéarité

Dans tout ce que suit, nous considérons le modèle linéaire multiple suivant 4.1 :

$$y_i = \sum_{j=1}^n a_j x_{ij} + \xi_i \quad (4.1)$$

où  $y_i$  est la variable forestière  $i$  à prédire (variable endogène),  $x_{ij}$  la  $j$ ième variable image explicative (variable exogène) et  $a_j$  son coefficient de régression associé et  $n$  le nombre de variables explicatives dans le modèle.

Il y a colinéarité entre une variable exogène et une autre (ou une combinaison linéaire d'autres exogènes) lorsque leur corrélation linéaire est élevée. Cette colinéarité est directement liée à la redondance de l'information, elle perturbe la lecture et la stabilité des résultats. Elle entraîne une exagération de la variance des estimateurs des coefficients de régression, les tests de Student ( $H_0$  sur les coefficients de régression) sont alors sous-estimés et les variables ne paraissent alors pas d'effet significatif. Les valeurs et signes des coefficients peuvent être contradictoires, ils ne concordent pas avec les connaissances du domaine. Enfin, les résultats sont peu robustes à l'ajout ou la suppression de quelques observations. La colinéarité peut fausser complètement l'interprétation des coefficients de la régression, il convient donc de la détecter afin de maîtriser ses effets (Guyon and Elissee, 2003).

Une approche simple consiste à observer les coefficients de détermination entre paires de variables exogènes, il y a alors colinéarité entre deux variables fortement corrélées (selon un seuil à définir). La colinéarité peut aussi être suspectée dès que le coefficient de détermination d'une paire de variables exogènes est comparable au coefficient de détermination multiple entre variable endogène et la paire de variables exogènes.

Ces méthodes ne détectant que la colinéarité bivariée, il faudrait effectuer la régression de chaque variable exogène avec les  $(p - 1)$  autres exogènes et étudier le coefficient de détermination associé afin d'observer la colinéarité multivariée. Mais il existe des tests de multicollinéarité dont le plus couramment rencontré est l'utilisation du facteur d'inflation de la variance (VIF).

$$VIF_j = \frac{1}{1 - R_j^2} \quad (4.2)$$

Avec  $R_j^2$  le coefficient de détermination obtenu par la régression de la  $j^{th}$  variable exogène avec toutes les autres. Une valeur de VIF existe donc pour chacune des variables prises en compte dans le modèle multiple. Plus le facteur d'inflation  $VIF_j$  est élevé, plus la variance de l'estimation associée  $V(\hat{a}_j)$  du coefficient de régression  $a_j$  sera forte. Cette estimation sera donc très instable et aura moins de chance d'être significative dans le test de nullité du coefficient dans la régression. Une règle de détection consiste alors à utiliser un seuil maximum acceptable pour cette valeur d'inflation de la variance. Autrement dit,  $VIF_j$  est égal à 1 quand il n'y a pas de colinéarité, les valeurs de  $VIF_j$  supérieures à 4 indiquent la présence de colinéarité qui nécessite une investigation plus poussée tandis que des valeurs de  $VIF_j$  supérieures à 10 indique une forte présence de colinéarité (Castillo et al., 2010; Haan, 2002; Kutner et al., 2004). Il existe aussi des tests plus rigoureux d'un point de vue statistique (exemple test de Farrar et Glauber), ces derniers s'appuyant sur une démarche équivalente : l'hypothèse nulle correspond à l'orthogonalité des variables, l'évaluation porte alors sur la mesure de l'écart à cette hypothèse.

Il est important de détecter la colinéarité, il faut ensuite pouvoir la traiter et la maîtriser. Les méthodes de sélection de variables permettent de palier ce problème.

## 4.2.2 Méthodes de sélection de variables

Il existe de nombreuses stratégies de sélection de variables, seules les plus courantes seront présentées ici. L'objectif est de déterminer un sous-ensemble de variables exogènes qui soient pertinentes et non redondantes pour expliquer la variable endogène  $y$  (Dodge and Rousson, 2004; Guyon and Elissee, 2003). La difficulté réside dans la détermination du bon nombre de variables pour la régression ainsi que la façon de choisir ces variables. La notion de parcimonie (préférer la simplicité, principe du rasoir d'Occam (Thorburn, 1918)) est ici prépondérante autant pour la lisibilité et l'interprétation du modèle que pour sa robustesse. Deux principales approches se distinguent : les sélections par optimisation d'un critère et les techniques basées sur le F partiel de Fisher.

### 4.2.2.1 Sélection par optimisation

Ce type de sélection propose de produire toutes les combinaisons possibles de variables exogènes afin de choisir celle qui maximise un critère de qualité. Ceci implique un grand nombre de cas à tester et donc un calcul coûteux. Les principaux critères de qualité sont les suivants :

1. Coefficient de détermination multiple  $R^2$  :

$$R^2 = 1 - \frac{SCR}{SCT} \quad (4.3)$$

où  $SCR = \sum(y_i - \hat{y}_i)^2$  est la somme des carrés résiduels et  $SCT = \sum(y_i - \bar{y}_i)^2$  est la somme des carrés totaux (ceux de la variable endogène).

Ce critère classique exprime la part de variance expliquée par le modèle. Le problème est que le coefficient augmente de façon mécanique avec l'ajout de variables (même si ces variables ne sont pas pertinentes), il ne convient donc pas pour comparer des solutions ayant des nombres de variables différents.

2.  $R^2$  ajusté :

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{(n-q-1)}}{\frac{SCT}{(n-1)}} = 1 - \frac{n-1}{n-q-1}(1 - R^2) \quad (4.4)$$

avec  $n$  le nombre de variables dans le modèle évalué.

En tenant compte du nombre de variables introduites dans le modèle (degrès de liberté), ce critère peut être utilisé pour comparer des solutions ayant

des nombres de variables différents. Cependant, en pratique, ce critère est trop permissif, il favorise les solutions avec un grand nombre de variables et par conséquent ne permet pas de respecter le principe de la parcimonie.

**3. AIC(Akaike) et BIC(Schwartz) :**

$$AIC = n \ln \frac{SCR}{n} + 2(q + 1) \quad (4.5)$$

$$BIC = n \ln \frac{SCR}{n} + \ln(n)(q + 1) \quad (4.6)$$

Ces critères permettent d'évaluer la précision du modèle en tenant compte de sa complexité (Burnham and Anderson, 2004). Il est à noter que le BIC pénalise plus fortement les modèles complexes à partir d'un certain nombre de variables ( 7 variables).

**4. Mallor's  $C_p$  :**

Pour un modèle avec  $p$  variables explicatives, cette statistique est définie par :

$$C_p = \left( \frac{SCR_p}{s^2} \right) - (n - 2p) \quad (4.7)$$

où  $s^2$  est le carré de l'erreur moyenne du modèle contenant toutes les variables explicatives,  $SCR_p$  est la somme des carrés des résidus du modèle contenant les  $p$  variables explicatives d'intérêt (Efron et al., 2004; Mallows, 1973). En général cette statistique est représentée en fonction du nombre de variables dans le modèle (à chaque étape de la sélection). Les modèles acceptables (dans le sens d'une minimisation du biais total pour les valeurs prédites) sont ceux pour lesquels la valeur de  $C_p$  se rapproche de  $p$ . L'observation de cette statistique permet d'estimer le bon nombre de variables à prendre en compte dans le modèle.

**5. PRESS(Predicted Residual Sum of Squares) :**

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i(-i))^2 \quad (4.8)$$

$\hat{y}_i(-i)$  est la prédiction de la valeur de la variable endogène pour l'observation

$i$  utilisée en donnée supplémentaire dans la régression numéro  $i$ . L'idée de ce critère est que pour avoir une estimation honnête des performances en prédiction, il ne faudrait pas que l'observation  $i$  participe à la construction du modèle quand on veut prédire sa valeur de l'endogène (Guyon and Elissee, 2003). Il est alors possible de définir une stratégie de sélection reposant uniquement sur ce critère sans tenir compte explicitement de la complexité. Si le  $R^2$  augmente avec l'ajout d'une variable, sans pouvoir explicatif, le PRESS va quant à lui se dégrader et signifier ainsi l'inutilité de cette variable.

Le critère PRESS est lié à la LOO RMSE (Leave-One-Out Root Mean Square Error) par la relation suivante :

$$LOORMSE = \frac{\sqrt{PRESS}}{n} \quad (4.9)$$

Ces sélections peuvent être ascendantes (Forward : ajouts successifs d'une variable exogène jusqu'à l'optimum) soit descendantes (Backward : retraits successifs jusqu'à l'optimum). Il faut ici noter que ces stratégies de sélection ne traitent pas directement le problème de colinéarité. Ceci est fait de manière implicite lorsque deux variables redondantes augmentent la complexité mais pas ou très peu le pouvoir explicatif du groupe de variables exogènes.

#### 4.2.2.2 Sélection basée sur le F partiel de Fisher

Le principe consiste à s'appuyer sur les propriétés inférentielles de la régression. Une variable est ajoutée si le carré du  $t$  de Student (qui suit une loi Fisher) indique que le coefficient de régression associé est significativement différent de zéro, sinon la variable est supprimée. La significativité dépend du risque  $\alpha$  fixé par l'opérateur; plus la valeur de  $\alpha$  est faible plus la sélection sera parcimonieuse. Comme précédemment, la sélection peut être ascendante ou descendante mais aussi hybride (Stepwise). La procédure Stepwise permet de vérifier que l'adjonction d'une variable ne provoque pas la suppression d'une variable déjà introduite et ainsi de suite jusqu'à convergence (plus d'ajout ni de retrait possible).

D'autres stratégies de sélection existent comme par exemple la sélection Stage-wise. Celle-ci consiste à chercher la variable exogène qui explique au mieux la part de variance non encore expliquée par la ou les exogène(s) sélectionnée(s) à l'étape précédente. La sélection par les corrélations partielles est aussi employée.

### 4.2.2.3 Sélection LARS (Least Angle Regression)

Least Angle Regression (LARS) est un algorithme récent de sélection de variables très utile et moins lourd que les traditionnelles méthodes de sélection Forward. Efron et al. (2004) en propose plusieurs variantes ; la méthode du LASSO, une méthode Stepwise ainsi qu'une méthode Forward stagewise. Comme pour les méthodes Forward classiques, l'initialisation se fait avec tous les coefficients de régression à 0 (ordonnée à l'origine présente ou pas), la première étape consiste à trouver le prédicteur le plus corrélé à la variable à estimer. Ensuite, le second prédicteur est choisi de façon à être le plus corrélé aux résidus courants. Puis, au lieu de continuer dans la direction du premier prédicteur, LARS va déterminer une direction équi-angulaire au deux prédicteurs retenus précédemment jusqu'à trouver le prédicteur le plus "corrélé" à cette direction. Enfin, LARS va continuer de cette façon, le long cette direction "least angle" pour un nombre d'étapes donné.

Il est à noter qu'il n'existe pas de meilleure stratégie générique et que ces différentes méthodes de sélection peuvent aboutir à des solutions différentes. Les questions principales étant la rigueur du critère de sélection utilisé, la conservation ou non de la variable choisie à une étape donnée et enfin le respect de la parcimonie et le contrôle de la colinéarité.

## 4.3 Méthodologie

### 4.3.1 Démarche proposée

Afin de préparer l'arrivée des images Pléiades, une première série d'expériences a été menée sur le site de Nezer en utilisant l'image QuickBird (Beguet et al., 2012). Cette première expérience a permis de tester et d'ajuster certains points méthodologiques qui ont ensuite été affinés à l'arrivée des images Pléiades en 2012 avec une seconde série d'expériences sur le site du Tagon (Beguet et al., 2014). La méthodologie présentée ainsi que les résultats obtenus rendent compte des deux séries d'expériences.

Un diagramme représentant la méthodologie proposée pour l'estimation des variables de structure forestière est présenté figure 4.1. Afin de pouvoir automatiser les processus, des échantillons images sont assemblés en mires, sont aux deux résolutions des images (panchromatique et multispectrale). Ils correspondent aux placettes où les mesures sont faites sur le terrain. Les attributs de

texture sont ensuite calculés de façon automatique sur ces échantillons d'image avec un grand nombre de paramètres spatiaux. Pour chaque échantillon image, un descripteur statistique (moyenne ou médiane) est conservé par attribut de texture calculé. Une sélection de variables automatique, basée sur les régressions linéaires (simples et multiples), permet ensuite de trouver les meilleurs attributs de texture décrivant chacune des variables de structure forestière. Pour les régressions linéaires multiples, différentes méthodes de sélection de variables classiques sont testées et une nouvelle méthode de sélection d'attributs, nommée Random PRiF, est proposée et sera détaillée dans ce qui suit.

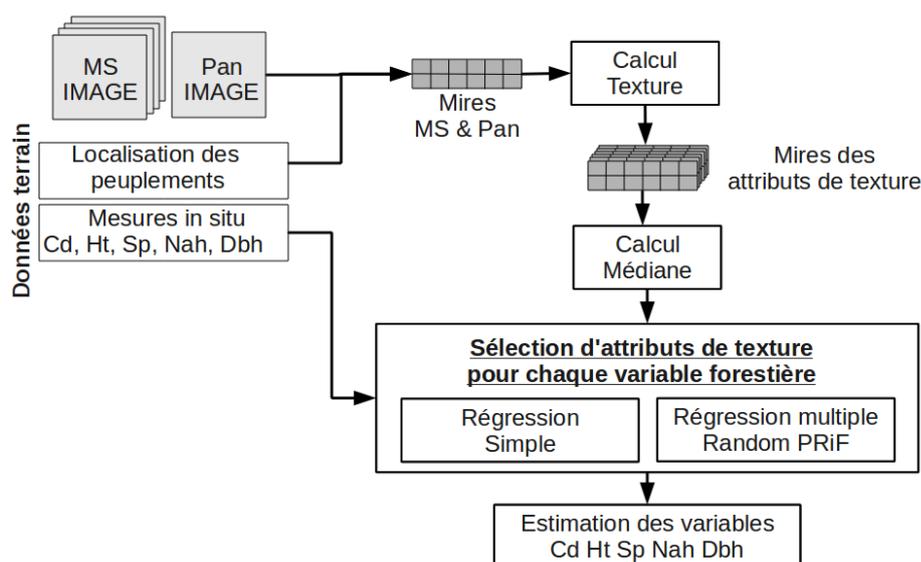


Figure 4.1 : Représentation schématique générale de la modélisation des variables forestières à partir des attributs image.

Enfin, les différentes méthodes de sélection de variables image seront évaluées ainsi que la performance des meilleurs modèles retenus et la précision des estimateurs des variables forestières.

### 4.3.2 Attributs de texture et paramètres

Dans notre approche, deux attributs de texture du premier ordre (moyenne et variance, appelés mean et var) seront calculés sur les images (la moyenne ne renseignant pas sur la texture mais sur la réflectance spectrale). Huit attributs de texture du second ordre (GLCM) ont été utilisés (Connors et al., 1984; Haralick et al., 1973) (cf. Table 2.1). Ces derniers sont considérés comme particulièrement

importants pour les applications en forêt (Castillo et al., 2010; Chehata et al., 2014; Coburn and Roberts, 2004; Franklin et al., 2001; Kayitakire et al., 2006; Murray et al., 2010; Ozdemir and Karnieli, 2011; Wood et al., 2012; Wunderle et al., 2007)

Les paramètres spatiaux des GLCM sont: le rayon de la fenêtre de calcul  $r$  (taille de la fenêtre =  $2r + 1$ ), le déplacement  $d$ , l'orientation  $o$  et sur le plan radiométrique le nombre de niveaux de quantification  $nbbin$ . L'un des inconvénients de ces attributs GLCM est le nombre de paramètres à fixer. Par conséquent un autre intérêt de l'automatisation du processus de sélection d'attributs est d'optimiser automatiquement les paramétrages spatiaux des attributs GLCM. Une gamme de paramètres limitée a été testée de manière à faire un compromis entre les coûts de calcul et les performances de prédiction attendues. Les gammes de  $r$  et  $d$  ont été fixées selon la résolution spatiale et sont présentées dans la table 4.1.

Tableau 4.1 : Gammes des valeurs des paramètres des GLCM.

	Rayon	Déplacement	Orientation
Donnée	$r$ (pixel)	$d$ (pixel)	$o$ (degrés)
Pan	5-25, pas: 5	1-10, pas :1	0 -135, pas: 45
MS	3-12, pas: 3	1-4, pas:1	0 -135, pas: 45

Le nombre de niveaux de quantification,  $nbbin$ , a été fixé à 8 comme souvent utilisé dans la littérature en télédétection et qui s'avère suffisant pour notre étude. Cependant ils sont définis sur la gamme de variation radiométrique rencontrée sur les portions d'image considérées.

Tous les attributs de texture ont été calculés à partir du panchromatique (Pan) et à partir des quatre bandes multispectrales (MS) : Bleu (B), vert (V), rouge (R), proche infrarouge (PIR), avec diverses combinaisons de valeurs de paramètres GLCM.

Lors de la première série d'expériences sur le site de Nezer, la texture a aussi été calculée sur l'image de NDVI (Normalized Difference Vegetation Index) ainsi que sur l'image d'intensité I (moyenne des quatre bandes spectrales). Ces choix se sont ensuite avérés non appropriés. Le calcul du NDVI nécessite de passer des comptes numériques originaux aux réflectances au sol, or les images utilisées

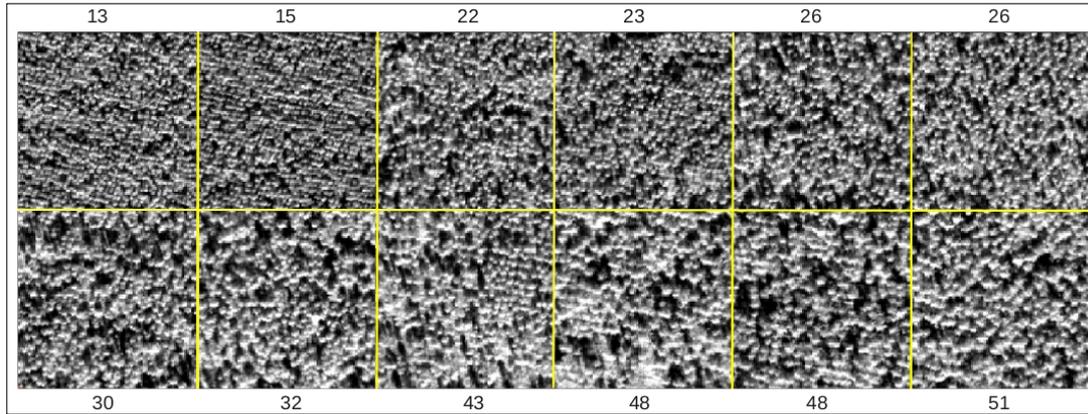
sont celles en comptes numériques. Le calcul de I aurait aussi mérité d'être fait au minimum sur les luminances au sommet de l'atmosphère (applications des coefficients de calibration aux comptes numériques). Certains résultats obtenus sur Nezer, à partir du NDVI et de I calculés sur les comptes numériques, seront toutefois présentés.

### 4.3.3 Échantillons d'image

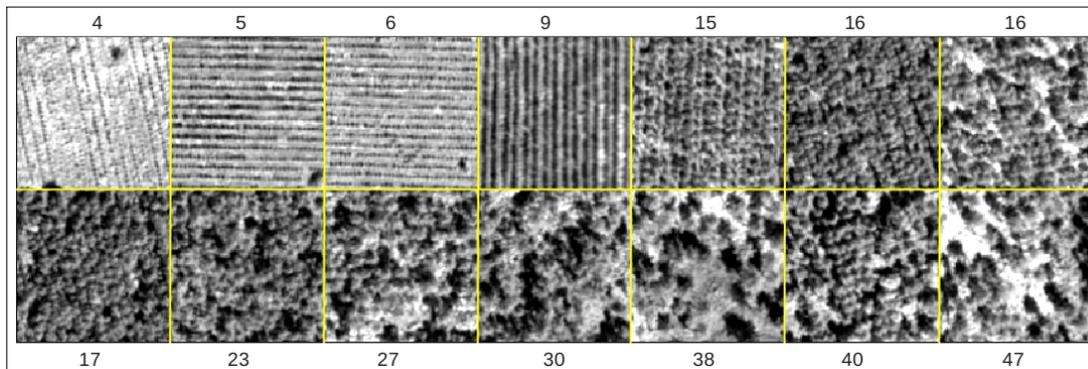
Le processus de sélection d'attributs de texture a été effectuée automatiquement sur des mires test regroupant de petits échantillons d'image de tous les peuplements où les variables forestières ont été mesurées in situ. En effet des milliers de descripteurs ont été calculés compte tenu du grand nombre de combinaisons possibles de valeurs de paramètres spatiaux.

Le processus de sélection de variables proposé évalue ainsi la capacité de multiples descripteurs à estimer les variables forestières avec des temps de calcul raisonnables. La figure 4.2 présente quelques échantillons d'image utilisés pour construire les mires test.

Chaque échantillon d'image couvre une zone carrée centrée sur une placette de mesure (*cf.* Section 2.2) avec une taille légèrement plus grande. Sa largeur est de 200 pixels pour le Pan et de 50 pixels pour le MS soit environ 120 mètres sur le site Nezer. Pour le site du Tagon, la largeur de l'échantillon est de 120 pixels en Pan et 30 pixels en MS soit environ 60 mètres.



(a)Site de Nezer (12 échantillons couvrant les 12 placettes de mesure faites sur le terrain, dimensions : 124\*124m)



(b)Site de Tagon (14 échantillons sur les 111 placettes de mesure faites sur le terrain, dimensions : 60\*60m)

*Figure 4.2 : Echantillons d'images panchromatiques pour les deux sites illustrant la variabilité des peuplements en termes de structure forestière. Les échantillons sont ordonnés par âge croissant.*

Les valeurs min et max de la radiométrie sur l'ensemble de la mire test ont été utilisés pour définir les niveaux de quantification des GLCM pour chaque bande spectrale afin de garantir une résolution radiométrique optimale des pixels de forêt sur le nombre de niveaux utilisé ( $nbbin = 8$ ).

La valeur médiane ou la moyenne de chaque échantillon d'image ont été ensuite calculées et utilisées pour dans l'étape suivante de sélection des attributs les plus pertinentes pour l'estimation des variables forestières. La valeur médiane a finalement été préférée à la valeur moyenne, qui est couramment utilisée en télédétection pour la modélisation, afin d'éviter de forts effets de lissage et réduire l'effet de possibles outliers après la comparaison faite en (*cf.* Section 4.4.1).

## 4.3.4 Sélection de variables par régressions linéaires

### 4.3.4.1 Régression linéaire simple

Les régressions linéaires ont été effectuées entre toutes les variables de structure forestière et dix attributs de texture, dérivés des bandes Pan et MS avec différents paramétrages. Le coefficient de détermination  $R^2$  a été utilisé comme critère d'identification du meilleur attribut de texture parmi l'ensemble des attributs issus de la bande Pan et parmi l'ensemble des attributs de texture issus des bandes MS pour chaque date et chaque site. Le traitement a été automatisé pour sélectionner l'attribut de texture menant au plus fort  $R^2$  pour chaque variable forestière. La précision de la modélisation linéaire a été estimée avec l'erreur quadratique moyenne (RMSE) calculée par validation croisée Leave-One-Out (nommée ici LOO RMSE).

### 4.3.4.2 Régression linéaire multiple

Ce type d'approche n'est pas nouveau. Ainsi Castillo et al. (2010) l'a appliqué en utilisant uniquement l'information multispectrale de SPOT-5 (bandes initiales, indices de végétation dérivés et attributs de texture GLCM). Wunderle et al. (2007) a utilisé la texture de l'image fusionnée (pan-sharpened) SPOT et une approche de régression basée sur une sélection d'attributs Stepwise pour modéliser un indice de structure du peuplement. Ozdemir and Karnieli (2011) ont modélisé les variables de structure forestière à partir d'attributs de texture issus des bandes multispectrales de WorldView2 à l'aide de régressions linéaires multiples pas à pas (step-by-step). Notre approche est plus raffinée car elle explore le potentiel de combiner les attributs d'Haralick provenant à la fois de la bande Pan et des bandes MS, et ainsi utilise conjointement différentes résolutions spectrales et spatiales. En outre, le traitement automatisé permet d'optimiser les paramètres spatiaux des GLCM tout en explorant diverses combinaisons d'attributs de texture permettant ainsi une modélisation multi-échelle. Les régressions multiples sont ainsi réalisées sur des milliers de descripteurs, un grand nombre d'entre eux étant susceptible d'être corrélés.

La principale difficulté en régression linéaire multiple est de minimiser la multicollinéarité au sein de l'ensemble d'attributs afin de générer des modèles stables et éviter l'overfitting. Pour cela nous avons utilisé le Facteur d'Inflation de la Variance (VIF), bon détecteur de multicollinéarité (*cf.* Section 4.2.1).

Dans un premier temps, une expérience (*cf.* Section 4.4.3.2.1) a été conduite afin de sélectionner, parmi trois méthodes de sélection de variable de la littérature, celle qui minimise la multicollinéarité dans le contexte de l'étude. La sélection classique Forward Fisher a été testée avec deux méthodes dérivées des Least Angular Regressions : le Lasso ainsi qu'une approche Stepwise. Cette première expérience a conduit à retenir la méthode LARS stepwise car réduisant le plus le problème de la colinéarité (*cf.* Section 4.4.3). Nous avons ensuite élaboré une nouvelle approche, nommée Random PRiF, et qui sera décrite dans la section suivante et comparée à la méthode LARS Stepwise et à une réduction de variables par ACP (Analyse en Composante Principale).

#### 4.3.4.3 Sélection par Random PRiF

Pour pallier les faiblesses des méthodes de sélection d'attributs traditionnelles, étape par étape, nous proposons une nouvelle méthode de sélection de sous-ensemble de variables, nommée "Random PRiF". Contrairement à la méthode LARS stepwise qui est une méthode incrémentale (en gardant la meilleure variable et en ajoutant progressivement de nouvelles variables complémentaires) et donc sous-optimale, notre méthode est globale et basée sur un schéma d'optimisation simple mais statistiquement robuste : échantillonnage aléatoire (avec remplacement) dans l'espace des variables. L'idée principale est d'évaluer, par un processus itératif, un grand nombre de sous-ensembles possibles d'attributs de texture choisis au hasard (avec remplacement) parmi l'ensemble des attributs, en minimisant un nouveau critère statistique qui combine de manière appropriée les critères PRESS et VIF. L'objectif est de trouver de manière automatique un sous-ensemble d'attributs ayant une performance prédictive forte tout en conservant une faible colinéarité entre les attributs du sous-ensemble. Le critère que nous avons nommé *PRiF* est défini comme suit :

$$PRiF = PRESS * \frac{\exp(V)}{V^2} \quad (4.10)$$

où  $V$  est le maximum du vecteur  $VIF$  associé aux  $m$  variables sélectionnées.

L'utilisation d'une fonction de  $VIF$  dans le  $PRiF$ , qui a un comportement hyperbolique, va favoriser des sous-ensembles avec une bonne performance de prédiction, tout en pénalisant ceux qui ont des valeurs élevées de  $VIF$  afin de garantir à la fois la précision de la prédiction et la stabilité dans les modèles résultants.

Comme on le verra, notre technique fournit des solutions de sous-ensembles

d'attributs optimaux au sein desquels l'équilibre entre les variables (importances relatives) est significativement meilleur que pour les sous-ensembles de variables obtenus par les autres méthodes testées. Certes, les approches step-by-step ou stepwise ont une plus faible complexité de calcul que notre approche mais elles présentent un risque élevé de converger vers un optimum local, conduisant ainsi à des modèles sous-optimaux. Pour Random PRiF, plus le nombre d'itérations est élevé, plus la prédiction sera performante, le sous-ensemble de variables tendant vers la solution optimale.

Cette approche innovante a été comparée à la méthode LARS stepwise ainsi qu'à l'Analyse en Composante Principale. L'ACP est une méthode de réduction de la dimensionalité (Manly, 1994), les composantes résultantes étant orthogonales, il est apparu intéressant de tester cette approche dans notre contexte, fort en colinéarité. L'ACP a été donc choisie comme méthode de référence sans multicollinéarité, ie.  $VIF = 1$  entre les composantes principales utilisées comme variables explicatives. Les résultats sont présentés Section 4.4.3.

## 4.4 Résultats

### 4.4.1 Choix préliminaires

Avant de développer une méthodologie pour l'estimation des variables de structure forestière, deux questions se sont posées. 1) Peut-on réduire la complexité des calculs en se contentant de calculer la texture sur l'image d'intensité (moyenne des bandes spectrales) plutôt que sur chacune des bandes spectrales ? 2) Une fois la texture calculée sur un échantillon d'image, quel descripteur statistique (représentant la texture de cet échantillon) conserver pour les régressions linéaires : moyenne ou médiane ?

Afin de pouvoir répondre à cette question, une expérience a été menée sur le site de Nezer ( $n=12$ ) en comparant les résultats des estimations (LOO RMSE) pour les variables *Cd*, *Ht* et *Sp* à partir des moyennes et des médianes de chaque échantillon d'image. La texture est calculée sur l'image d'intensité (I), sur chacune des bandes multispectrales et sur la bande panchromatique. Les résultats sont présentés table 4.2.

Variable	I		MS		Pan	
	Moy	Med	Moy	Med	Moy	Med
Cd(m)	0.369	0.357	0.4	0.302	0.326	<b>0.246</b>
Ht(m)	1.483	1.484	<b>1.401</b>	1.45	1.899	1.539
Sp(m)	0.457	0.464	0.369	0.37	0.415	<b>0.363</b>

Tableau 4.2 : Les attributs de texture I, MS et Pan les plus performants pour l'estimation des variables forestières. Site de Nezer, image Quickbird d'Octobre.

Ces résultats indiquent que le fait d'utiliser la moyenne ou la médiane d'un échantillon d'image comme valeur représentative fournit des résultats comparables avec tout de même un avantage pour la médiane. De plus, la médiane étant moins sensible aux distributions asymétriques et aux outliers, elle a un effet de lissage moins important et par conséquent elle sera préférée à la moyenne pour la suite de cette étude. La texture calculée sur l'image d'intensité fournit des résultats moins performants que la texture calculée sur chacune des bandes spectrales prises séparément, elle ne semble donc pas être une alternative satisfaisante.

#### 4.4.2 Solutions par régression simple

Les régressions simples ont été appliquées séparément sur les ensembles d'attributs Pan et MS pour les deux sites. En raison des larges gammes de valeurs des paramètres spatiaux, 1730 et 2368 attributs de texture ont été testés pour Pan et MS respectivement. Pour chaque variable forestière, de nombreuses relations linéaires significatives avec des attributs de texture ont été obtenues (considérant une pvalue < 0,01 pour  $R^2$ ).

Les attributs de texture les plus performants, avec les plus forts  $R^2$ , sont présentés dans les tables 4.3 et 4.4 pour Pan et MS sur les sites de Nezer et du Tagon respectivement. Pour chaque variable forestière, les meilleurs attributs Pan et MS sont présentés avec leur paramétrage correspondant ainsi que la bande spectrale.

##### 4.4.2.1 Les descripteurs retenus

Pour le site de Nezer (Table 4.3), toutes les variables forestières étudiées sont bien décrites par un attribut de texture avec des  $R^2$  variant de 0,89 à 0,98.

Tableau 4.3 : Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Nezer, image d'Octobre,  $n=12$ .

var. Forêt	Bande	Attribut	$r$	$d$	$0$	$R^2$	pvalue	LOO RMSE
Cd	B	cluspro	3	1	0	0.970	$< 10e - 06$	0.30
	Pan	inertia	10	5	45	0.979	$< 10e - 06$	0.25
Sp	V	inertia	3	2	0	0.961	$< 10e - 06$	0.37
	Pan	inertia	5	1	45	0.944	$< 10e - 06$	0.40
Ht	V	inertia	6	2	0	0.946	$< 10e - 06$	1.5
	Pan	inertia	25	1	45	0.905	$< 10e - 06$	1.6
Nah	B	corr	3	1	45	0.891	$< 10e - 06$	140
	Pan	inertia	5	1	135	0.909	$< 10e - 06$	37
Dbh	V	inertia	12	1	45	0.938	$< 10e - 06$	0.029
	Pan	inertia	10	5	45	0.927	$< 10e - 06$	0.029

Les performances sont pour la plupart similaires entre Pan et MS sur cet ensemble réduit de peuplements ( $n = 12$ ). Les attributs de texture calculés sur la bande verte semblent être les plus efficaces pour le jeu de données MS. De plus, les attributs de texture de second ordre sont nettement plus représentatifs de la structure forestière que les attributs du premier ordre. L'inertie est le descripteur d'Haralick le plus fréquemment retenu, cela pour les deux résolutions et avec différents paramétrages. Il est intéressant de remarquer que pour des attributs de texture Pan, le paramètre de rayon de la fenêtre varie de 5 à 25 pixels en fonction de la variable forestière recherchée. Cela confirme qu'aucun a priori ne peut être fait sur les paramètres de GLCM, ce qui justifie une recherche de grande envergure et automatique des valeurs optimales de paramètres pour l'estimation des variables de structure forestière.

Pour le site du Tagon, les régressions simples ont été traitées pour les trois dates: Juin 2012, Août 2012 et Février 2013. La qualité des prévisions est inférieure à celle observée pour Nezer ce qui était prévisible étant donnée la forte diversité de la structure forestière sur le site.  $R^2$  varie de 0,32 (Nah) à 0,73 (Sp) en Juin, de 0,34 (Nah) à 0,78 (Sp) en Août et de 0.36 (Nah) à 0.84 (Ht) en février.

De même que pour le site de Nezer, la bande verte offre les meilleurs modèles simples (pour le jeu d'attributs MS) en Juin et en Août. En février, la bande PIR apparaît comme la plus informative. Pour les trois dates, les attributs de texture MS semblent être plus pertinents pour l'estimation des variables de structure forestière que les attributs de texture Pan (/cf. les valeurs de  $R^2$  et LOO RMSE). La variance de la bande verte, attribut de premier ordre, donne les meilleurs

résultats pour les images d'été. Contrairement à Nezer, où les rangées d'arbres ont toutes la même orientation, le site du Tagon se caractérise par une grande variété d'orientations des rangs. La variance, qui est insensible à l'anisotropie de la texture de l'image, semble être plus robuste aux variations d'orientation que les attributs du second ordre.

Tableau 4.4 : Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image de Juin,  $n=111$ .

var. Forêt	Bande	Attribut	$r$	$d$	$0$	$R^2$	pvalue	LOO RMSE
Cd	V	var	6			0.650	$< 10e - 06$	1.397
	Pan	corr	10	2	45	0.505	$< 10e - 06$	1.659
Sp	V	var	3			0.737	$< 10e - 06$	0.979
	Pan	var	25			0.572	$< 10e - 06$	1.242
Ht	PIR	invdif	12	3	135	0.635	$< 10e - 06$	4.450
	Pan	corr	15	3	45	0.550	$< 10e - 06$	4.993
Nah	PIR	corr	12	1	135	0.445	$< 10e - 06$	682.112
	Pan	corr	10	2	45	0.321	$< 10e - 06$	716.776
Dbh	V	var	6			0.650	$< 10e - 06$	0.086
	Pan	corr	10	2	45	0.533	$< 10e - 06$	0.099

Tableau 4.5 : Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image d'Août,  $n=111$ .

var. Forêt	Bande	Attribut	$r$	$d$	$0$	$R^2$	pvalue	LOO RMSE
Cd	V	var	3			0.699	$< 10e - 06$	1.30
	Pan	cluspro	25	3	45	0.517	$< 10e - 06$	1.67
Sp	V	var	3			0.785	$< 10e - 06$	0.89
	Pan	cluspro	25	3	45	0.604	$< 10e - 06$	1.22
Ht	V	var	3			0.640	$< 10e - 06$	4.4
	Pan	corr	25	3	45	0.521	$< 10e - 06$	5.2
Nah	R	var	3			0.386	$< 10e - 06$	677
	Pan	corr	25	3	45	0.343	$< 10e - 06$	704
Dbh	V	var	3			0.687	$< 10e - 06$	0.081
	Pan	cluspro	25	3	45	0.512	$< 10e - 06$	0.103

Tableau 4.6 : Les attributs de texture MS et Pan les plus performants pour l'estimation des variables de structure forestière. Site de Tagon, image de Février,  $n=111$ .

var. Forêt	Bande	Attribut	$r$	$d$	$0$	$R^2$	pvalue	LOO RMSE
Cd	PIR	var	12			0.768	$< 10e - 06$	1.136
	Pan	cluspro	25	3	45	0.649	$< 10e - 06$	1.404
Sp	PIR	cluspro	12	4	135	0.703	$< 10e - 06$	1.033
	Pan	cluspro	25	3	45	0.666	$< 10e - 06$	1.100
Ht	PIR	entropy	12	4	135	0.837	$< 10e - 06$	2.981
	Pan	cluspro	15	3	45	0.640	$< 10e - 06$	4.471
Nah	PIR	inertia	12	4	0	0.360	$< 10e - 06$	690.371
	Pan	invdif	25	1	45	0.366	$< 10e - 06$	706.976
Dbh	PIR	inertia	12	4	0	0.807	$< 10e - 06$	0.064
	Pan	cluspro	20	3	45	0.663	$< 10e - 06$	0.085

Pour l'image de février, les résultats sont significativement différents des résultats obtenus pour les deux images d'été. Ce sont des attributs du second ordre calculés sur la bande PIR qui donnent les meilleurs résultats, ceci indique bien une sensibilité de la texture à la saison (variation de la contribution du sous-bois, variation de l'angle solaire) cependant il est délicat d'interpréter l'influence de ces différents facteurs sur cette texture.

Des valeurs similaires de  $R^2$  et RMSE sont observées pour les images de juin et août, bien que légèrement supérieures pour l'image août. En effet, la texture d'image des mêmes peuplements varie probablement entre les deux dates en raison de petites différences de phénologie, mais aussi les angles solaires et les angles de visée. Par exemple, l'hétérogénéité spatiale de la réflectance du sous-bois était probablement légèrement plus petite en août en raison du stade phénologique (la croissance des feuilles a pris fin en août pour toutes les espèces contrairement à juin). La position du soleil induit de légèrement plus grandes ombres projetées sur le sol en août, un angle de vue zénithale inférieur en août accentue aussi ce phénomène. En dépit de ces petites différences, la texture de l'image est bien représentative de la structure de la forêt pour les deux dates d'été. Les performances des meilleurs sous-ensembles d'attributs d'image sont relativement bonnes et similaires pour les deux dates, la méthodologie semble ainsi être robuste aux conditions d'acquisition d'images.

Les performances de l'image d'hiver sont supérieures. On peut suspecter un effet de l'état de la végétation du sous bois en hiver (surface foliaire verte quasi nulle

(Yauschew-Raguenees, 2012) entraînant une variabilité plus faible de la réflectance du sous bois, donc une contribution plus importante de la surface arborée dans la texture) et de longueur des ombres des couronnes d'arbres plus élevée en hiver.

#### 4.4.2.2 Les performances de prédiction

Pour les deux sites et les différentes dates, les meilleures performances de prédiction (LOO RMSE) ont été obtenues pour l'espacement des arbres ( $Sp$ ), le diamètre des couronnes ( $Cd$ ) et la hauteur des arbres ( $Ht$ ) qui sont les variables de la forêt qui influent directement sur la texture de l'image. Toutefois, le diamètre des troncs  $Dbh$ , qui n'a aucun effet direct sur la texture, a également montré de bonnes valeurs de  $R^2$ , des valeurs proches de celles obtenues pour  $Cd$ . Cela vient des fortes relations allométriques linéaires entre ces deux variables (cf. Figure 3.5) déjà démontré pour le pin maritime par (Porté et al., 2000) par exemple. La densité de plantation  $Nah$  a montré les plus faibles corrélations linéaires avec les attributs de texture contrairement à  $Sp$  qui est calculé comme une transformation non linéaire de  $Nah$  (cf Eq. 4.8).

Les figures 4.3, 4.4, 4.5 et 4.6 comparent les valeurs mesurées sur le terrain et les valeurs prédites par le meilleur modèle de régression simple pour les cinq variables de structure forestière considérées. Aucun comportement aberrant, excepté un ou deux points en février (ce dont on ignore la cause) n'a été observé sur les résidus. L'hypothèse de relations linéaires pour les variables forestières est clairement confirmée pour le site de Nezer, mais le petit nombre d'observations ne permet pas une interprétation fiable des résultats.

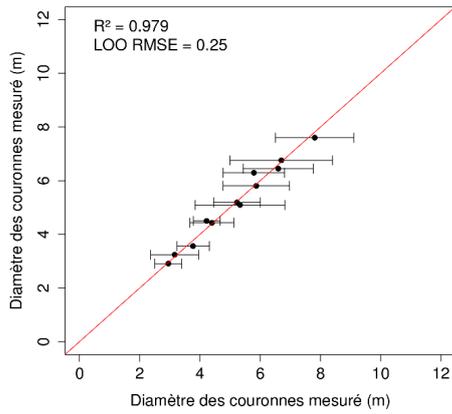
Le site du Tagon est plus instructif car il fournit une plus grande variabilité en termes de structure forestière. Tout d'abord, la variance (voir les écarts-types de la figure 4.4) des variables mesurées sur le terrain doit être prise en considération lors de l'interprétation de la qualité des prévisions. Un comportement hétéroscédastique est observé, plus prononcé pour le diamètre des couronnes  $Cd$  que pour la hauteur des arbres  $Ht$ . Cela met en évidence le fait que plus les peuplements sont âgés, plus la variance intra-peuplement de  $Cd$  est forte. En effet, avec une augmentation de l'âge et de l'espacement entre les arbres (éclaircies), la répartition spatiale des arbres devient plus hétérogène, induisant ainsi des variations locales de la croissance des couronnes (en raison de la concurrence entre les arbres). Aussi, plus les arbres sont âgés, plus il est difficile de mesurer

avec précision le diamètre des couronnes sur le terrain, l'incertitude sur la mesure est donc plus élevée.

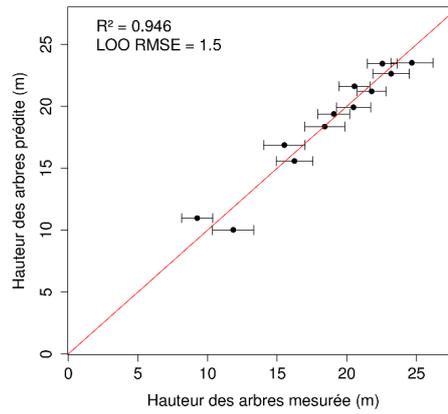
Deuxièmement, les relations entre les valeurs mesurées et prédites ne sont pas complètement linéaires à l'exception des  $Sp$  (*cf.* Figure 4.4). Les valeurs prédites pour  $Cd$  et  $Ht$  ont été surestimées pour les jeunes peuplements (*cf.* Figure 4.4). Par exemple, pour les hauteurs mesurées sur le terrain  $Ht < 5\text{m}$ , les valeurs prédites varient de 5 à 10 m. En outre, pour les trois variables forestières, la variance résiduelle apparaît d'autant plus grande que la variabilité des variables mesurées sur le terrain est élevée. Ainsi, l'utilisation d'un modèle de régression simple est insuffisante pour l'estimation de variables forestières.

Ces résultats préliminaires ont montré le potentiel d'une sélection automatique des attributs de texture les plus informatifs (à savoir le meilleur descripteur d'Haralick et son paramétrage optimal) pour estimer les variables de structure forestière. Cependant, l'utilisation d'un seul descripteur de l'information texturale ne paraît pas suffisant pour décrire la complexité de la structure forestière. La section suivante propose donc de déterminer des prédictors multi variables, combinant des attributs de texture multi-résolutions (issus des deux résolutions) et multi-échelles (avec différents paramétrages spatiaux), ceci afin d'exploiter au mieux l'information texturale pour estimer les variables de la structure forestière.

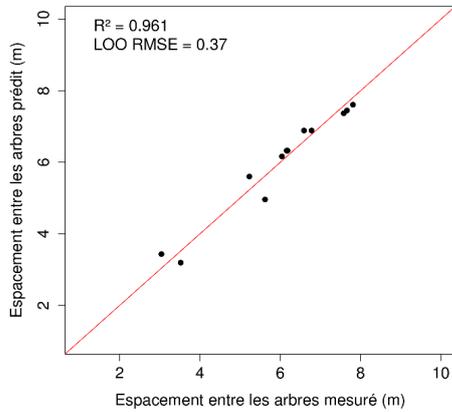
4. ESTIMATION DES VARIABLES FORESTIÈRES À PARTIR DE LA TEXTURE THR  
PAR MODÉLISATION STATISTIQUE



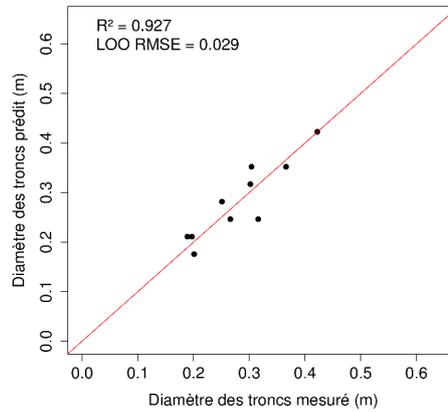
(a) Diamètre des couronnes (Cd)



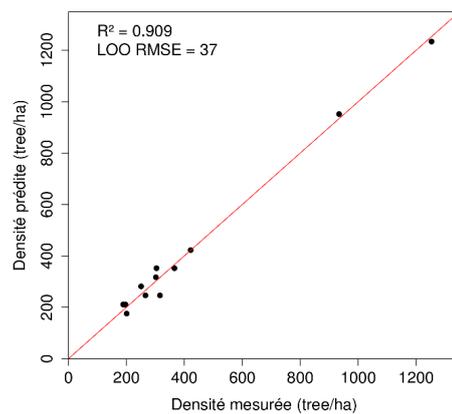
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)

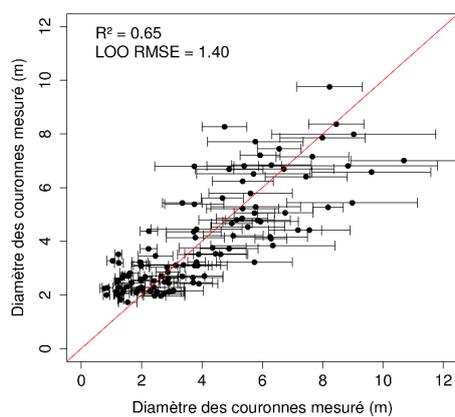


(d) Diamètre des troncs (Dbh)

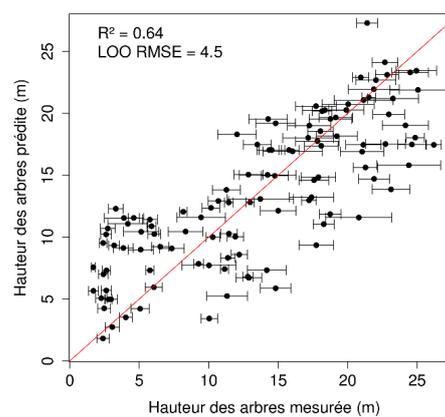


(e) Densité (Nah)

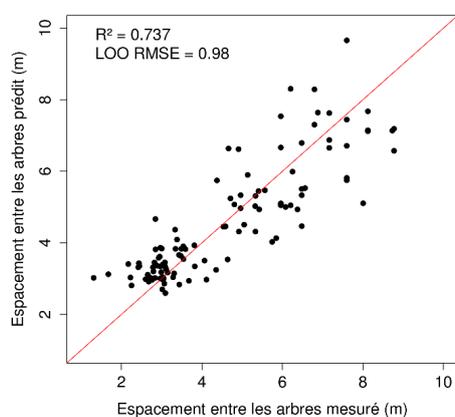
Figure 4.3 : Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site de Nezer. Les barres correspondent à  $\pm 1$  écart type.



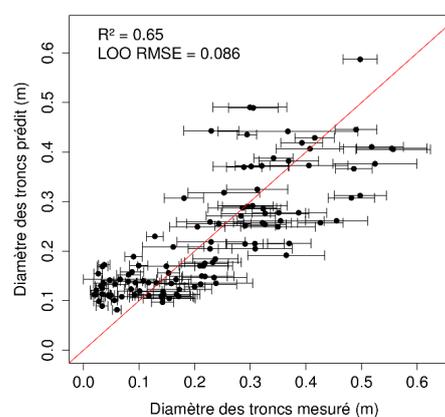
(a) Diamètre des couronnes (Cd)



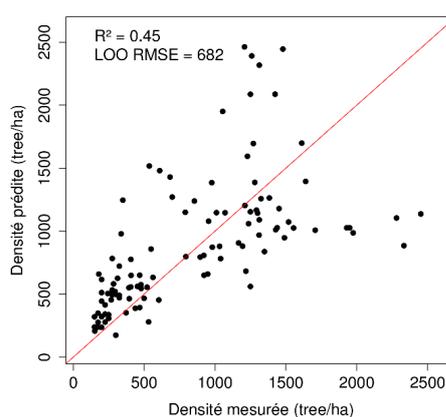
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)



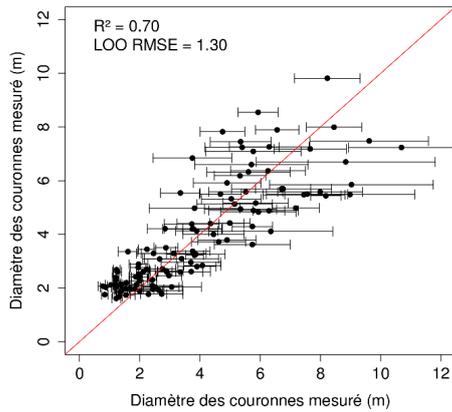
(d) Diamètre des troncs (Dbh)



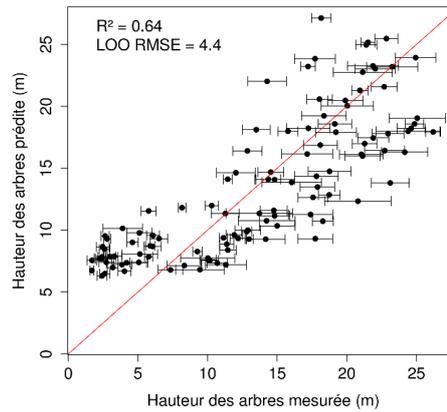
(e) Densité (Nah)

Figure 4.4 : Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Juin. Les barres correspondent à  $\pm 1$  écart type.

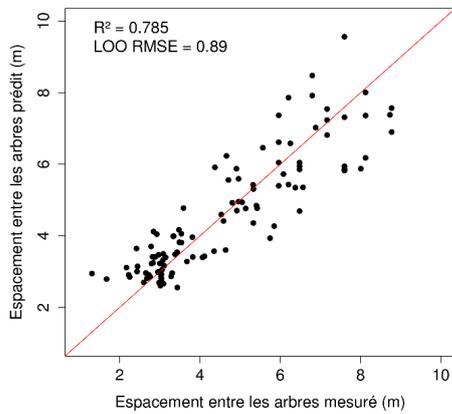
#### 4. ESTIMATION DES VARIABLES FORESTIÈRES À PARTIR DE LA TEXTURE THR PAR MODÉLISATION STATISTIQUE



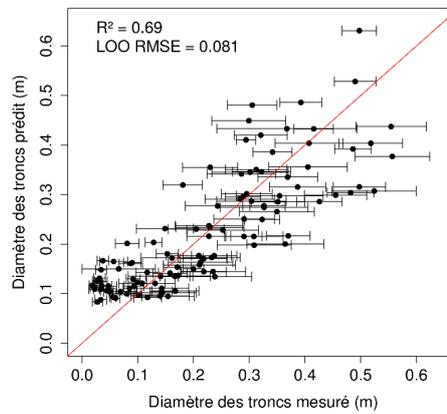
(a) Diamètre des couronnes (Cd)



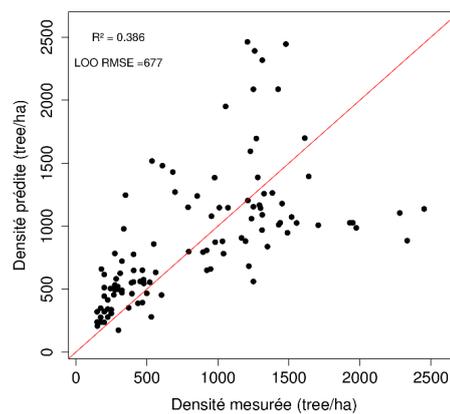
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)

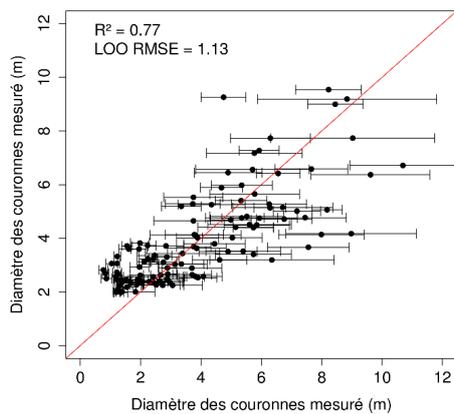


(d) Diamètre des troncs (Dbh)

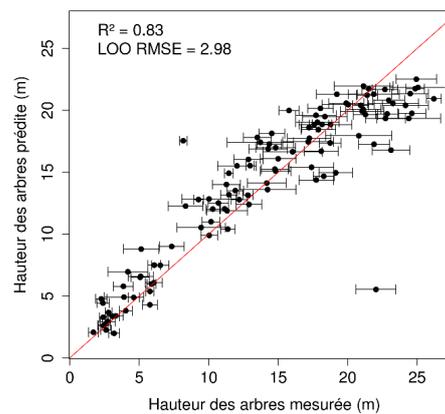


(e) Densité (Nah)

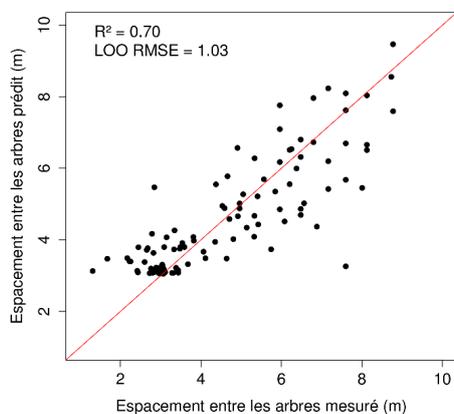
Figure 4.5 : variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Aout. Les barres correspondent à  $\pm 1$  écart type.



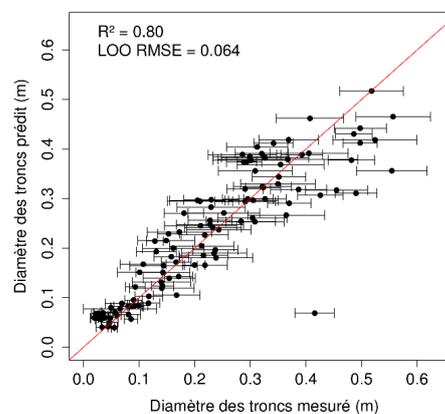
(a) Diamètre des couronnes (Cd)



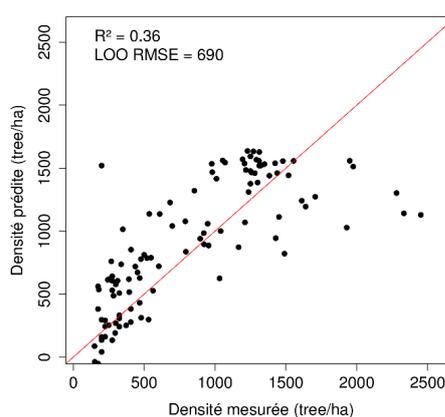
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)



(d) Diamètre des troncs (Dbh)



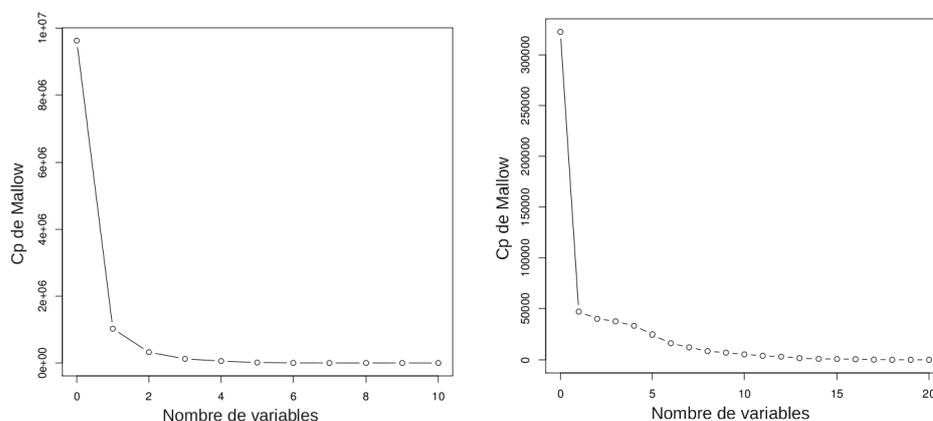
(e) Densité (Nah)

Figure 4.6 : Variables de structure forestière mesurées vs. estimations dérivées des meilleurs modèles simples, site du Tagon, image de Février. Les barres correspondent à  $\pm 1$  écart type.

### 4.4.3 Solutions par régression multiple

#### 4.4.3.1 Nombre de variables

La construction de modèles de régression multiple pose le problème du choix du nombre de variables à prendre en compte dans le modèle. De façon mécanique, plus on ajoute de variables dans le modèle, plus la précision de ce dernier augmente. Cependant, il n'est pas cohérent d'utiliser un trop grand nombre de variables, en effet, plus un modèle est complexe moins il est stable. Ce principe de parcimonie devant être respecté, il faut donc préférer un modèle avec le moins de variables possibles. Cela étant dit, il n'existe pas de méthode universelle pour déterminer le nombre optimal de variables dans un modèle de régression multiple. Différents essais ont été conduits pour déterminer de manière automatique le nombre de variables optimal, aucun n'a été satisfaisant. La statistique  $C_p$  de Mallows (*cf.* 4.2.2) est généralement utilisée pour déterminer le bon nombre de variables à prendre en compte (Efron et al., 2004; Mallows, 1973), cette étape nécessite cependant de déterminer un seuil en observant la courbe d'évolution de  $C_p$  en fonction du nombre de variables. Dans notre cas, nous avons choisi de fixer, par site, un nombre de variables identique pour chacune des variables forestières considérées et pour pour toutes les méthodes de sélection de variables testées.



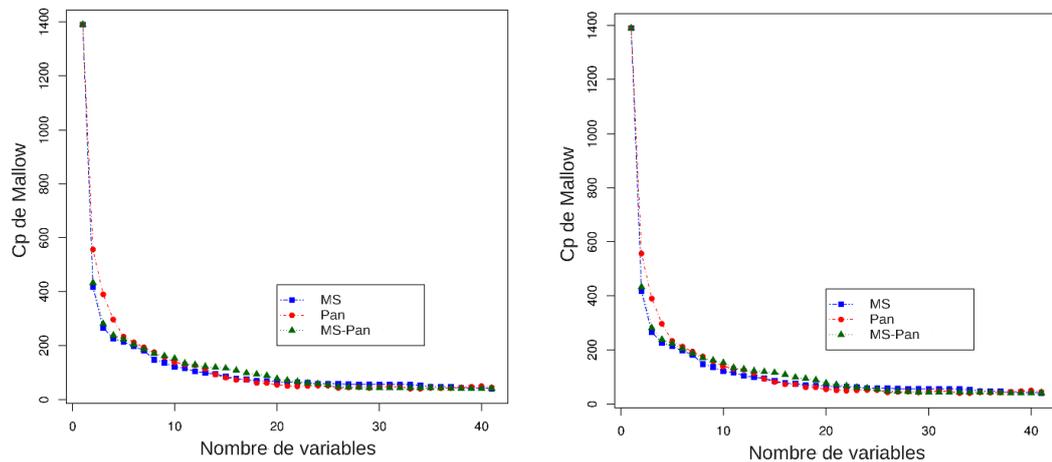
(a) Estimation du diamètre des couronnes, LARS stepwise, n=12  
(b) Estimation de l'âge des peuplements, LARS stepwise, n=184

Figure 4.7 : Evolution de la statistique  $C_p$  de Mallows en fonction du nombre d'attributs de texture pour l'estimation du diamètre des couronnes et de l'âge des peuplements pour le site de Nezer (image d'octobre).

Pour le site de Nezer, la figure 4.7 nous a conduit à fixer le nombre de variables à 3 pour l'estimation des variables forestières (n=12). Pour l'estimation

de l'âge des peuplements ( $n=184$ ), étant donné que le nombre d'observations est nettement plus important, nous avons fixé à 10 le nombre de variables dans le modèle. Ces courbes ont été obtenues en utilisant la méthode de sélection LARS stepwise appliquée à l'ensemble des attributs de texture calculés sur les 4 bandes spectrales et la bande panchromatique (MS-Pan).

Pour le site du Tagon, l'observation des courbes d'évolution de la statistique  $C_p$  de Mallow 4.8 est plus délicate. On observe que le palier est atteint pour 20 variables lorsque l'on estime  $C_d$  avec la méthode LARS stepwise, cela quelque soit l'ensemble d'attributs de texture considéré et pour les deux dates d'été. Les courbes correspondant aux autres variables forestières ont la même allure, par conséquent, elles ne sont pas présentées. Cependant, ce seuil de nombre de variables de 20 est très élevé et cette statistique apparaît non adaptée. Un autre indicateur a alors été utilisé et qui mesure l'importance relative des variables (la statistique LMG définie par (Gromping, 2006) a été utilisée). Le tableau 4.11 indique l'importance relative de chacune des variables dans le modèle à 5 variables. A partir de là, étant donné la très faible importance de la dernière variable, nous avons décidé de fixer le nombre de variables à 5, seuil à partir duquel l'évolution de  $C_p$  est ralentie également. Ce nombre permet d'avoir un compromis entre parcimonie et efficacité.



(a) Estimation du diamètre des couronnes, LARS stepwise, image de Juin  
(b) Estimation du diamètre des couronnes, LARS stepwise, image d'Aout

Figure 4.8 : Evolution de la statistique  $C_p$  de Mallow en fonction du nombre d'attributs de texture pour l'estimation du diamètre des couronnes pour le site du Tagon en Juin et Aout à partir des trois ensembles de données MS, Pan et MS-Pan.

Pour la suite de cette étude, le nombre d'attributs de texture pour les modèles de régressions multiples a été donc fixé à 3 pour le site de Nezer et à 5 pour le site du Tagon pour toutes les variables forestières autres que l'âge.

#### 4.4.3.2 Comparaison des méthodes de sélection de variables

**4.4.3.2.1 Expérience Nezer** Dans un premier temps, les sélections de variables Forward, Lasso et LARS stepwise ont été testées sur le site de Nezer ( $n=12$ ) pour l'ensemble des variables forestières. Les ensembles d'attributs de texture Pan et MS ont été traités séparément. Il faut noter qu'au moment de cette expérience, le calcul de la texture se faisait aussi sur le NDVI (Normalized Difference Vegetation Index).

Les résultats sont présentés pour l'estimation de l'âge des peuplements au tableau 4.7 pour le panchromatique et au tableau 4.8 pour le multispectral. Les comportements observés étant similaires pour les autres variables, ils ne sont pas présentés. Pour chaque méthode de sélection de variables, les trois attributs de texture retenus sont présentés ainsi qu'une mesure de l'importance relative de chacune des variables (LMG), le VIF pour observer la colinéarité et le critère PRESS pour observer la performance de prédiction.

Les tableaux 4.7 et 4.8 nous permettent d'observer le comportement de différentes méthodes de sélection de variables face au problème de la colinéarité. On peut observer que, si la méthode classique Forward permet (dans le cas du multispectral) d'obtenir les meilleures performances en termes de PRESS, les valeurs de VIF obtenues sont critiques ( $>4$ ) (Castillo et al., 2010; Haan, 2002; Kutner et al., 2004). On peut en déduire que cette méthode ne permet pas d'éviter le problème de colinéarité. La méthode du Lasso n'est pas satisfaisante en termes de performance de prédiction ni en termes de colinéarité. La méthode LARS stepwise quant à elle, fournit pour les deux jeux de données, des sous-ensembles d'attributs de texture pertinents en termes de prédiction tout en minimisant la colinéarité (VIF très proche de 1). Cette méthode apparaît alors comme étant la plus satisfaisante pour notre problématique de recherche de sous-ensemble d'attributs dans un contexte de grande dimension avec beaucoup de redondances (et donc de colinéarité).

Tableau 4.7 : Comparaison des performances des méthodes de sélection de variables pour l'estimation de l'âge, données panchromatiques ( $n = 12$ ). Nezer

Methodes	Bande	Attribut	$r$	$d$	$0$	moment	Lmg	VIF	PRESS
Forward	Pan	invdif	25	6	45	M	0.521	4.534	38.113
	Pan	clushade	25	5	90	M	0.393	2.779	
	Pan	energy	10	3	0	M	0.084	2.247	
Lasso	Pan	invdif	25	6	45	M	0.356	4.653	105.213
	Pan	inertia	20	8	0	V	0.320	6.247	
	Pan	clushade	20	6	90	M	0.322	3.725	
LARS Stepwise	Pan	invdif	25	6	45	M	0.906	1.041	80.423
	Pan	inertia	25	3	45	M	0.078	1.069	
	Pan	corr	25	2	0	V	0.015	1.102	

Tableau 4.8 : Comparaison des performances des méthodes de sélection de variables pour l'estimation de l'âge, données multispectrales ( $n = 12$ ). Nezer

Methodes	Bande	Attribut	$r$	$d$	$0$	moment	Lmg	VIF	PRESS
Forward	G	variance	12			M	0.434	7.738	116.77
	G	clushade	3	1	0	V	0.393	5.299	
	PIR	clushade	6	2	90	M	0.351	3.794	
Lasso	G	variance	12			M	0.336	5.365	361.283
	NDVI	clushade	3	2	45	V	0.332	5.612	
	R	corr	6	2	0	V	0.330	5.082	
LARS Stepwise	G	variance	12			M	0.819	1.118	115.731
	PIR	corr	6	2	45	M	0.038	1.039	
	R	corr	9	4	0	V	0.142	1.081	

L'observation de la statistique LMG (mesure d'importance relative de la variable dans le modèle) est intéressante, on peut observer que si la méthode LARS stepwise évite la colinéarité, elle propose des ensembles déséquilibrés, la première variable étant largement plus influente que les suivantes. La méthode Lasso, quant à elle, propose des solutions très équilibrées, c'est aussi le cas pour la sélection Forward. Il semble alors que l'équilibre entre variables et colinéarité soit lié. On peut alors se demander s'il est possible de trouver des sous-ensembles de variables équilibrés et exempt de colinéarité. C'est cette idée qui a alors motivé la création d'une nouvelle approche de sélection de variable, Random PRiF, dont l'objectif est à la fois d'éviter la colinéarité, d'équilibrer les importances des variables (ceci afin de pouvoir profiter au maximum de la complémentarité des

attributs de texture) tout en maintenant une bonne performance de prédiction.

La méthode LARS stepwise a été ensuite appliquée à l'estimation de l'âge des peuplements sur l'ensemble du site de Nezer. Dix variables issues de l'ensemble de données complet (MS-Pan) ont été utilisées. La LOO RMSE obtenue est de 5,25 ans soit à peu près 10% de la dynamique de la variable âge sur l'ensemble du site. La figure 4.9 montre que si les valeurs d'âge estimées sont satisfaisantes pour un âge inférieur à 30 ans, elles ne le sont plus ensuite. Ceci indique notamment le fait que l'âge est une variable trop synthétique pour décrire la complexité de la structure forestière, celle-ci n'évolue plus linéairement avec l'âge et après 30 ans elle peut montrer une très forte variabilité.

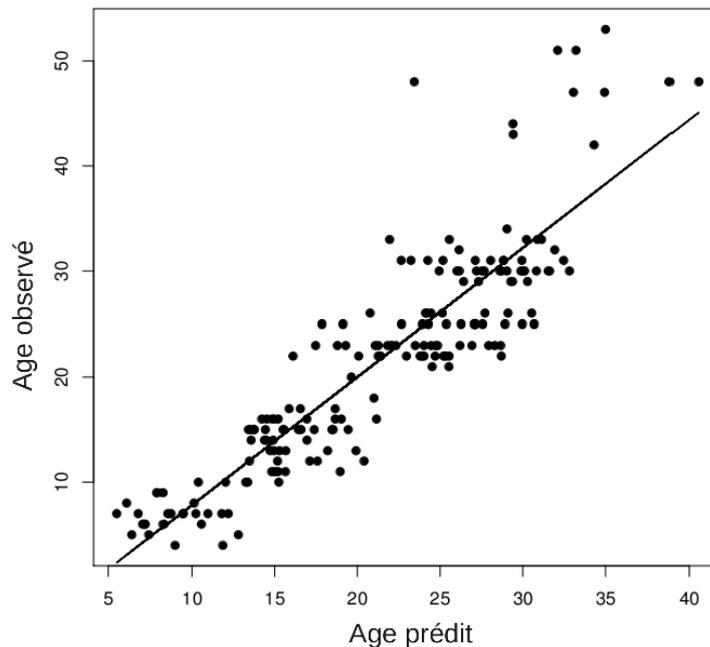


Figure 4.9 : Estimation de l'âge des peuplements, LARS stepwise, site de Nezer ( $n=184$ ). LOO RMSE = 5.25 ans.

Cette première expérience a permis d'appréhender notre problème sur un cas simple (Site très homogène, données peu nombreuses et de grande qualité), il reste à tester l'approche sur un cas plus complexe: le site du Tagon. La méthode LARS stepwise et la réduction de la dimensionnalité par Analyse en Composantes Principales ont été appliquées en comparaison avec Random PRiF dans une expérience plus complète.

**4.4.3.2 Expérience globale** Les trois méthodes de sélection de variables concurrentes (ACP, LARS stepwise et Random PRiF) ont été comparées sur les deux sites, pour chacune des variables forestières et pour trois ensembles d'attributs de texture: Pan seul, MS seul et MS-Pan qui contient l'ensemble des attributs de texture calculés aux deux résolutions. Ainsi, 30 expériences différentes (2 sites \* 3 jeux de données \* 5 variables forestières) comparant les trois méthodes de sélection de variables ont été effectuées. Les tableaux 4.9 et 4.10 illustrent les résultats en montrant les performances ( $R^2$ , PRESS, VIF, LOO RMSE) des sous-ensembles d'attributs obtenus par chacune des méthodes testées sur l'ensemble de données MS-Pan du site de Nezer et sur l'ensemble de données Pan du site du Tagon (les autres tables correspondantes sont données en Annexe B).

Une procédure de vote a été appliquée sur les 30 expériences pour identifier la méthode de sélection d'attributs la plus pertinente pour prédire les variables forestières en utilisant comme critère la plus petite LOO RMSE. Notre méthode de sélection d'attributs Random PRiF obtient la meilleure performance de prédiction dans près de 77% des cas, suivie par LARS stepwise. L'ACP est la méthode la moins satisfaisante dans tous les cas (*cf.* Tableaux 4.9 et 4.10). Lorsque LARS stepwise surpasse Random PRiF, les LOO RMSE correspondantes sont très proches. Ces résultats confirment la pertinence et l'efficacité du critère PRiF introduit dans un contexte global de recherche aléatoire.

En outre, la colinéarité, qui est un problème critique dans la régression linéaire multiple, est négligeable avec des valeurs maximales de  $VIF$  inférieures au seuil critique établi de 4 (Castillo et al., 2010; Haan, 2002; Kutner et al., 2004) pour toutes les solutions (à la fois pour LARS stepwise et Random PRiF) (*cf.* tableaux 4.9 et 4.10). Quant à l'ACP, qui construit un nouvel espace de variables composé de variables non corrélées (d'importance décroissante) et assure, par définition, aucune colinéarité, le  $VIF$  n'est pas un critère pertinent ( $VIF = 1$  pour toutes les variables d'un sous-ensemble donné). Les résultats montrent que notre méthode Random PRiF a un bon comportement en termes de stabilité et de fiabilité du modèle, présentant des valeurs maximales de  $VIF$  inférieures à 4 dans tous les cas. LARS stepwise surpasse généralement Random PRiF par rapport à la minimisation du  $VIF$  tandis que c'est le contraire en ce qui concerne les critères de qualité de prédiction (PRESS et LOO RMSE). C'est l'essence même de notre nouveau critère de sélection des variables qui a été construit de manière à accroître la précision des prévisions tout en gardant une colinéarité raisonnablement basse. Nous

4. ESTIMATION DES VARIABLES FORESTIÈRES À PARTIR DE LA TEXTURE THR  
PAR MODÉLISATION STATISTIQUE

Tableau 4.9 : Analyse comparative des méthodes de sélection d'attributs : critères statistiques des meilleurs sous ensembles de trois attributs (MS-Pan), image d'Octobre, site de Nezer.

var. Forêt	Méthode de sélection	R <sup>2</sup> multiple	PRESS	Max(VIF)	LOO RMSE
Cd	PCA	0.917	5.458	1.000	0.67
	LARS stepwise	0.995	0.190	1.288	0.13
	Random PRiF	0.996	0.168	2.421	0.12
Sp	PCA	0.906	6.149	1.000	0.72
	LARS stepwise	0.994	0.535	1.025	0.21
	Random PRiF	0.996	0.233	2.734	0.14
Ht	PCA	0.903	66.034	1.000	2.3
	LARS stepwise	0.992	5.114	1.070	0.7
	Random PRiF	0.994	2.791	2.519	0.5
Nah	PCA	0.871	464568	1.000	197
	LARS stepwise	0.998	8154	1.287	26
	Random PRiF	0.997	5582	3.404	22
Dbh	PCA	0.892	0.029	1.000	0.049
	LARS stepwise	0.987	0.003	1.226	0.016
	Random PRiF	0.993	0.001	1.270	0.009

Tableau 4.10 : Analyse comparative des méthodes de sélection d'attributs : critères statistiques des meilleurs sous ensembles de cinq attributs (Pan), image d'Août, site de Tagon.

var. Forêt	Méthode de sélection	R <sup>2</sup> multiple	PRESS	Max(VIF)	LOO RMSE
Cd	PCA	0.534	331.114	1.000	1.73
	LARS stepwise	0.752	180.304	1.357	1.27
	Random PRiF	0.800	129.015	2.293	1.08
Sp	PCA	0.548	214.820	1.000	1.39
	LARS stepwise	0.765	116.219	2.001	1.02
	Random PRiF	0.808	84.263	1.860	0.87
Ht	PCA	0.556	2904.690	1.000	5.1
	LARS stepwise	0.837	1144.369	2.052	3.2
	Random PRiF	0.851	1032.486	1.666	3.1
Nah	PCA	0.225	66281517	1.000	773
	LARS stepwise	0.523	51588535	1.408	682
	Random PRiF	0.500	46197877	1.748	645
Dbh	PCA	0.541	1231.149	1.000	0.105
	LARS stepwise	0.818	475.571	3.817	0.065
	Random PRiF	0.823	461.019	2.132	0.064

favorisons délibérément le PRESS plus que le  $VIF$  puisque le premier critère a un impact plus important sur les performances des modèles globaux multilinéaires. L'ACP fournit évidemment les plus bas  $VIF$  possibles, mais cela n'a jamais permis à cette méthode bien connue de dépasser en terme de prédiction les deux autres méthodes évaluées dans tous les 30 cas testés.

Le tableau 4.11 souligne l'importance de chaque attribut de texture dans les sous-ensembles sélectionnés par LARS stepwise et Random PRiF pour les sites de Nezer et du Tagon dans les cas les plus démonstratifs (les autres cas sont en Annexe B).

Tableau 4.11 : Importance de variables pour les sous-ensembles d'attributs sélectionnés (*MS-Pan*) par LARS stepwise et Random PRiF, pour les sites de Nezer (*image d'Octobre*) et du Tagon (*image d'Août*).

var. Forêt	Nezer		Tagon	
	LARS Stepwise	Random PRiF	LARS Stepwise	Random PRiF
Cd	0.921	0.752	0.761	0.788
	0.011	0.272	0.126	0.141
	0.066	0.220	0.069	0.040
			0.037	0.024
			0.005	0.005
Sp	0.959	0.445	0.811	0.613
	0.035	0.442	0.124	0.180
	0.004	0.111	0.051	0.148
			0.005	0.029
			0.006	0.028
Ht	0.975	0.372	0.674	0.599
	0.018	0.326	0.154	0.234
	0.006	0.300	0.142	0.074
			0.019	0.061
			0.008	0.029

La méthode Random PRiF conduit à des valeurs d'importance maximale généralement plus faibles que LARS stepwise qui présente une très grande préférence pour la première variable. Par exemple pour Nezer, pour la variable *Ht*, Random PRiF fournit un sous-ensemble équilibré où les trois caractéristiques de texture sont à peu près de même importance tandis que LARS stepwise conduit à un sous-ensemble d'attribut complètement déséquilibré. En effet, cette dernière méthode

est basée sur un schéma d'optimisation incrémental qui n'est pas la meilleure façon de combiner les caractéristiques dans un cadre de régression multiple. Bien que menant à des solutions sous-optimales, cette méthode étape par étape a un avantage sur notre méthode: une complexité de calcul inférieure. Ce comportement équilibré de la méthode Random PRiF a été clairement observé sur le site de Nezer ainsi que sur le site Tagon bien que moins spectaculaire. Ainsi, l'approche d'optimisation aléatoire globale utilisée dans Random PRiF permet d'optimiser la complémentarité entre les attributs.

Enfin, les résultats des différentes expériences mentionnées précédemment ont été résumés par variable forestière et par site. Pour chaque variable forestière, sont conservés le meilleur ensemble de données de départ et la méthode de sélection de variable correspondante, ces résultats sont présentés dans le tableau 4.12. On peut noter que, l'ensemble des données multirésolution MS-Pan conduit généralement à de meilleurs modèles. Cela met en évidence le potentiel de combiner les deux résolutions spatiales et spectrales. En ce qui concerne les performances des méthodes de sélection, obtenues sur les 30 expériences, Random PRiF apparaît comme la meilleure alternative. En conséquence, les sections suivantes se concentrent sur les solutions fournies par Random PRiF à partir de l'ensemble de données MS-Pan.

*Tableau 4.12 : Meilleures solutions multiples pour les sites de Nezer et du Tagon (Août) avec la méthode de sélection d'attribut retenue et l'ensemble de données respectif.*

var. Forêt	Site	Donnée	Méthode de sélection
Cd	Nezer	MS-Pan	Random PRiF
	Tagon	Pan	Random PRiF
Sp	Nezer	Pan	Random PRiF
	Tagon	MS-Pan	LARS Stepwise
Ht	Nezer	MS	Random PRiF
	Tagon	MS-Pan	LARS Stepwise
Nah	Nezer	MS-Pan	Random PRiF
	Tagon	MS-Pan	Random PRiF
Dbh	Nezer	MS-Pan	Random PRiF
	Tagon	MS-Pan	LARS Stepwise

### 4.4.3.3 Apport des solutions multiples

Les sous-ensembles d'attributs de texture les plus performants obtenus par Random PriF à partir de l'ensemble de données MS-Pan sont présentés dans les tableaux 4.13, 4.14, 4.15, 4.16 pour les sites de Nezer et du Tagon (pour l'ensemble des dates).

On peut remarquer que la régression linéaire multiple combine différents attributs de texture de second ordre avec différents paramétrages et surmonte ainsi le problème d'anisotropie (dû à la distribution spatiale des arbres en lignes). Par exemple, le meilleur sous-ensemble d'attributs pour estimer *Nah* sur le site Nezer combine deux résolutions et trois orientations différentes (45 , 90 et 135 ). Un autre exemple est le meilleur sous-ensemble d'attributs pour estimer *Cd* sur le site du Tagon (avec l'image d'août), il est composé de deux moments de GLCM Pan avec la même taille de la fenêtre ( $r = 15$  pixels) et des déplacements différents ( $d = 1$  et  $d = 5$  pixels) permettant ainsi une analyse multi-échelles. Trois attributs de texture de résolution inférieure, calculés sur la bande PIR, avec différents paramétrages, complètent le sous-ensemble.

Tableau 4.13 : Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Nezer.

var. Forêt	Bande	Attribut	$r$	$d$	$\theta$	multiple $R^2$	VIF	LOO RMSE
Cd	B	cluspro	3	1	0	0.996	2.421	0.12
	Pan	corr	25	9	135		2.382	
	B	corr	9	2	0		1.039	
Sp	Pan	invdif	5	2	135	0.996	2.735	0.14
	Pan	clushade	20	7	135		1.998	
	V	cluspro	6	4	0		2.412	
Ht	R	corr	9	3	0	0.994	1.164	0.5
	R	cluspro	6	2	0		2.346	
	R	inertia	3	1	90		2.519	
Nah	PIR	corr	12	2	45	0.997	1.069	22
	R	energy	3	2	90		3.404	
	Pan	inertia	15	2	135		3.389	
Dbh	Pan	invdif	10	4	45	0.993	1.270	0.009
	R	corr	3	2	0		1.062	
	V	entropy	6	3	135		1.282	

Tableau 4.14 : Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image de Juin.

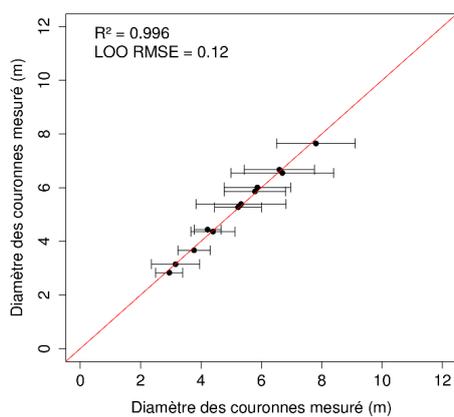
var. Forêt	Bande	Attribut	$r$	$d$	$\theta$	multiple $R^2$	VIF	LOO RMSE
Cd	B	corr	12	1	45	0.786	1.2631	1.12
	Pan	inertia	25	1	0		1.1171	
	V	var	6				1.0401	
	R	entropy	6	1	135		1.5407	
	Pan	harcorr	15	6	45		1.7416	
Sp	V	var	6			0.775	1.5934	0.923
	PIR	corr	12	2	135		1.6326	
	B	entropy	12	2	135		2.0378	
	B	clushade	6	2	90		2.0032	
	Pan	inertia	10	3	45		1.1189	
Ht	PIR	corr	12	2	135	0.786	2.949	3.6
	Pan	harcorr	20	1	90		1.8302	
	B	cluspro	12	2	90		1.1809	
	Pan	inertia	10	5	0		1.5226	
	V	var	6				1.7206	
Nah	Pan	clushade	15	4	135	0.541	1.9918	639.712
	PIR	corr	12	2	135		1.3308	
	Pan	harcorr	25	1	90		1.6796	
	R	corr	3	1	0		2.2297	
	B	invdif	6	1	135		2.124	
Dbh	B	harcorr	9	4	90	0.827	1.183	0.062
	PIR	harcorr	9	1	45		1.5488	
	B	entropy	9	2	135		1.1495	
	V	var	6				1.6013	
	Pan	energy	20	2	0		1.3255	

Tableau 4.15 : Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image d'Août.

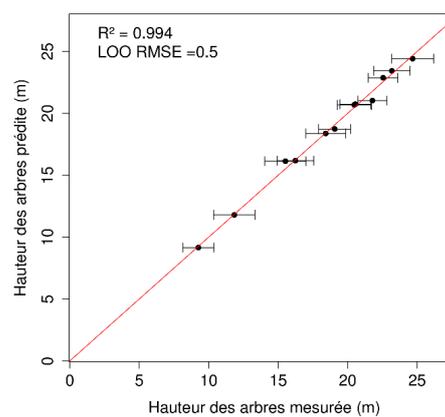
var. Forêt	Bande	Attribut	$r$	$d$	$\theta$	multiple $R^2$	VIF	LOO RMSE
Cd	Pan	harcorr	15	1	135	0.807	2.261	1.10
	Pan	cluspro	15	5	0		2.240	
	PIR	harcorr	12	4	0		2.140	
	PIR	clushade	3	2	45		1.009	
	PIR	inertia	12	3	135		2.128	
Sp	Pan	harcorr	20	8	0	0.819	1.127	0.84
	B	harcorr	9	2	0		1.918	
	R	cluspro	12	1	45		2.23	
	R	var	6				2.303	
	V	harcorr	3	1	45		2.100	
Ht	R	cluspro	9	1	0	0.831	1.607	3.1
	Pan	entropy	25	4	0		1.778	
	Pan	inertia	25	1	0		1.256	
	Pan	clushade	15	10	135		1.148	
	Pan	inertia	25	2	90		1.290	
Nah	PIR	harcorr	9	2	0	0.625	1.562	586
	V	clushade	3	1	0		1.107	
	V	inertia	6	5	0		2.106	
	PIR	corr	3	1	135		1.225	
	PIR	invdif	12	5	0		2.526	
Dbh	Pan	inertia	20	2	0	0.808	1.055	0.066
	V	var	6				1.192	
	V	corr	12	1	135		1.188	
	B	harcorr	12	1	90		2.021	
	V	harcorr	9	5	0		2.070	

Tableau 4.16 : Solutions multiples combinant données panchromatiques et multispectrales obtenues par Random PRiF, site de Tagon, image de Février.

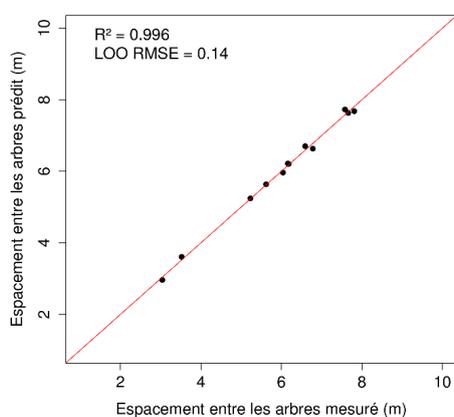
var. Forêt	Bande	Attribut	$r$	$d$	$0$	multiple $R^2$	VIF	LOO RMSE
Cd	Pan	entropy	10	8	0	0.827	2.326	1.101
	V	clushade	12	4	135		1.102	
	Pan	cluspro	10	9	45		1.612	
	Pan	entropy	5	2	45		1.831	
	Pan	cluspro	25	4	0		2.231	
Sp	Pan	cor	5	2	90	0.822	1.477	0.841
	Pan	clushade	20	2	90		1.768	
	Pan	harcorr	15	7	45		1.365	
	Pan	corr	5	2	0		1.634	
	Pan	invdif	10	7	0		1.390	
Ht	R	invdif	9	4	45	0.894	1.862	2.50
	Pan	cluspro	15	2	0		2.232	
	Pan	invdif	20	2	90		1.597	
	Pan	cor	15	2	0		1.322	
	Pan	invdif	25	7	90		1.781	
Nah	Pan	clushade	15	1	135	0.478	1.268	657
	Pan	corr	25	2	90		1.569	
	Pan	corr	15	2	0		1.540	
	V	clushade	12	4	0		1.202	
	Pan	harcorr	15	8	45		1.853	
Dbh	R	harcorr	9	3	0	0.859	1.293	0.058
	Pan	corr	15	4	90		1.620	
	Pan	invdif	10	1	135		1.258	
	Pan	corr	25	5	0		1.586	
	Pan	cluspro	20	3	45		1.919	



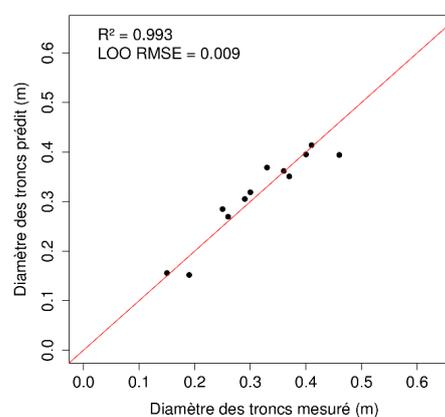
(a) Diamètre des couronnes (Cd)



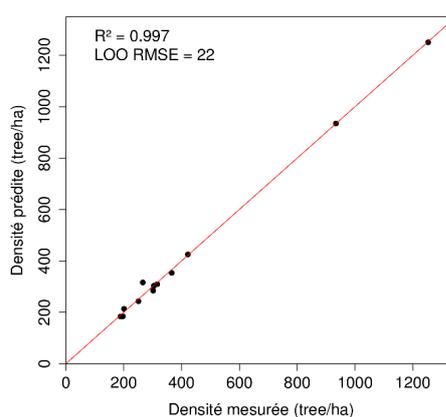
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)



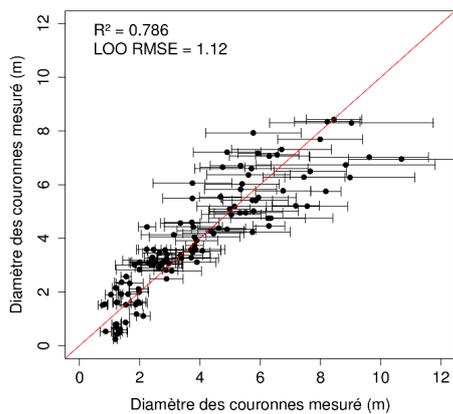
(d) Diamètre des troncs (Dbh)



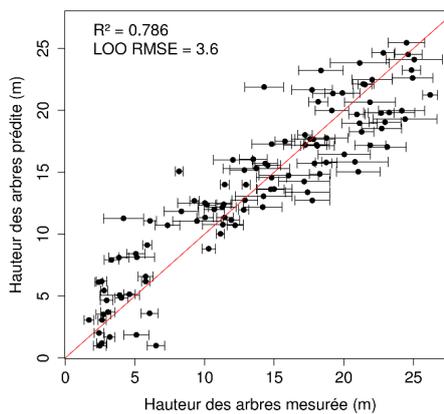
(e) Densité (Nah)

Figure 4.10 : Variables de structure forestière mesurées vs. estimations prédites par des meilleurs modèles multiples, site de Nezer. Les barres correspondent à  $\pm 1$  écart type.

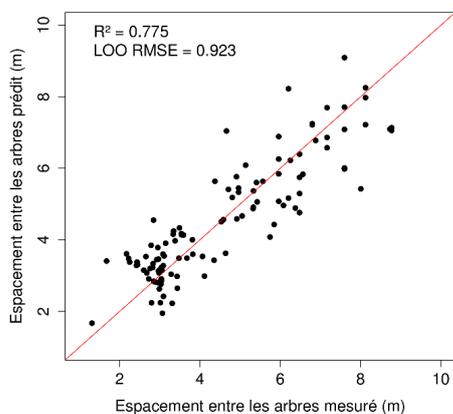
4. ESTIMATION DES VARIABLES FORESTIÈRES À PARTIR DE LA TEXTURE THR  
PAR MODÉLISATION STATISTIQUE



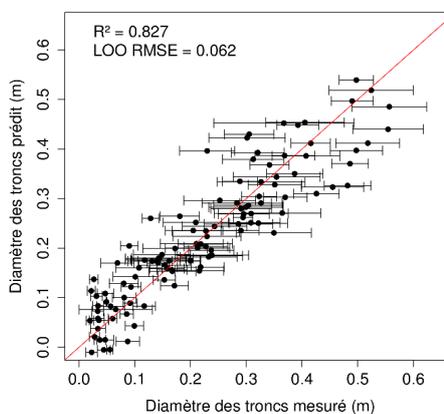
(a) Diamètre des couronnes (Cd)



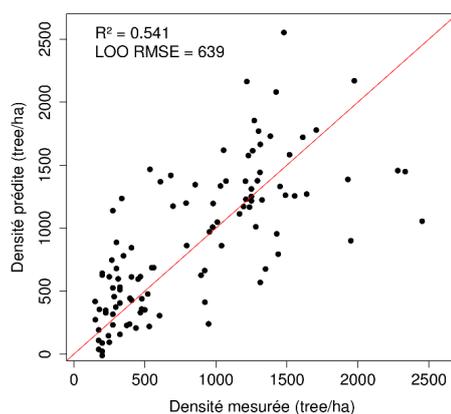
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)

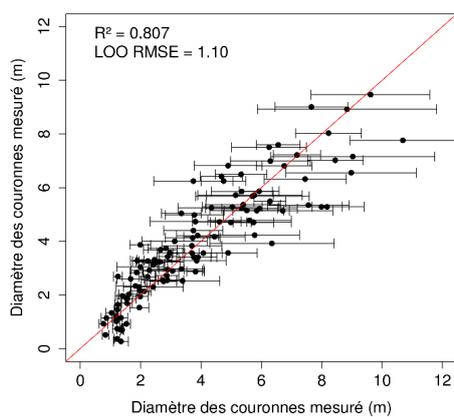


(d) Diamètre des troncs (Dbh)

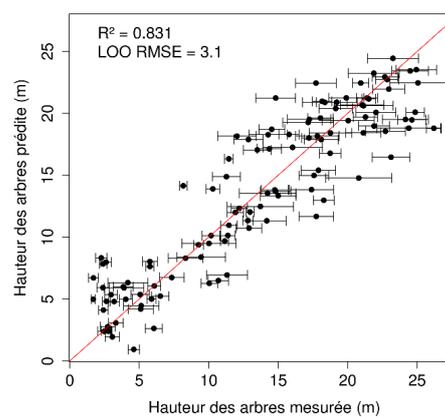


(e) Densité (Nah)

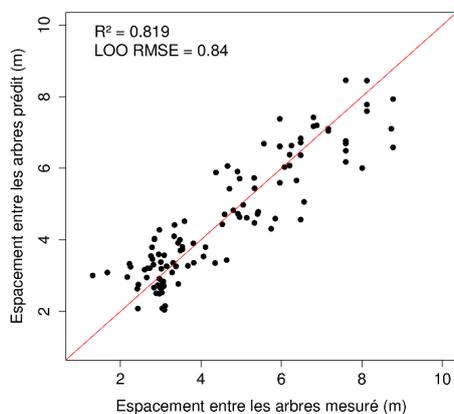
Figure 4.11 : variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Juin. Les barres correspondent à  $\pm 1$  écart type.



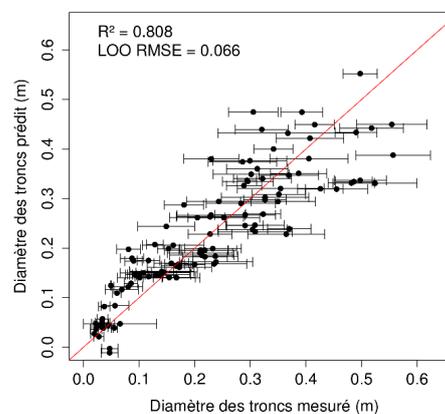
(a) Diamètre des couronnes (Cd)



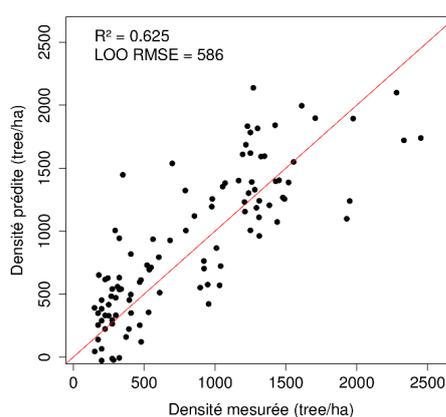
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)



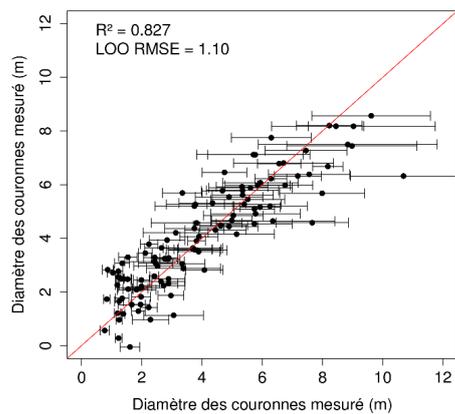
(d) Diamètre des troncs (Dbh)



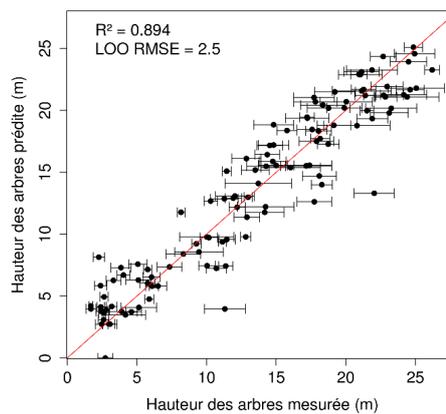
(e) Densité (Nah)

Figure 4.12 : variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Aout. Les barres correspondent à  $\pm 1$  écart type.

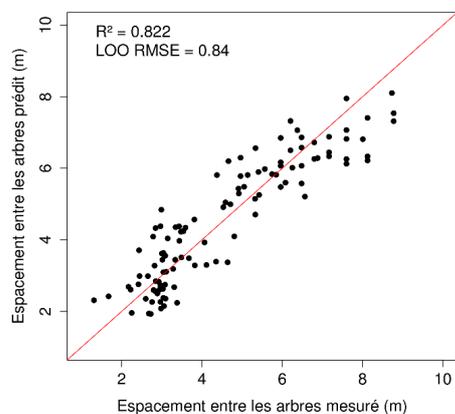
4. ESTIMATION DES VARIABLES FORESTIÈRES À PARTIR DE LA TEXTURE THR  
PAR MODÉLISATION STATISTIQUE



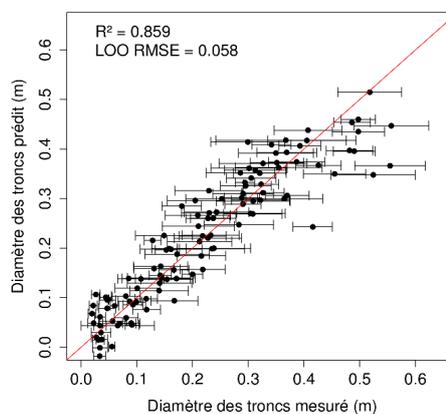
(a) Diamètre des couronnes (Cd)



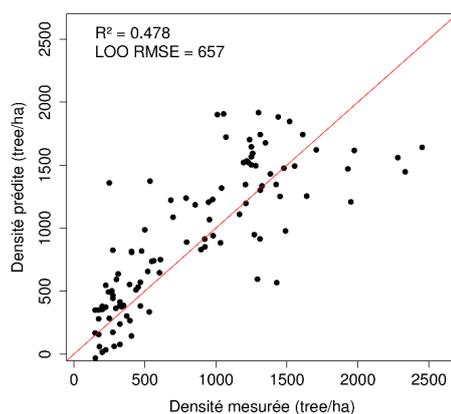
(b) Hauteur des arbres (Ht)



(c) Espacement entre les arbres (Sp)



(d) Diamètre des troncs (Dbh)



(e) Densité (Nah)

Figure 4.13 : variables de structure forestière mesurées vs. estimations prédites des meilleurs modèles multiples, site du Tagon, image de Février. Les barres correspondent à  $\pm 1$  écart type.

Les figures 4.10, 4.11, 4.12, 4.13 donnent une comparaison entre les valeurs mesurées in situ et leurs estimations prédites à partir des meilleurs modèles de régression multiples pour l'ensemble des variables forestières. Les modèles multiples réduisent clairement les résidus en comparaison avec les modèles simples (*cf.* Figure 4.3, 4.4, 4.5, 4.6) pour les deux sites. Les résidus sont également distribués de façon plutôt aléatoire tout autour de la ligne 1:1; reflétant une variance résiduelle à peu près constante quelle que soit la valeur de la variable forestière à la différence des solutions simples.

Le tableau 4.17 résume les performances globales des solutions simples et multiples pour l'ensemble des données disponibles. Les performances sont comparées en terme de LOO RMSE. Tous les scores de prédiction sont améliorés avec les solutions multiples.

Tableau 4.17 : Comparaison des performances des solutions simples et multiples (MS-Pan) pour les deux sites.

var. Forêt	Site	LOO RMSE		Gamme de variations
		Single	Multiple	Observée in-situ
Cd (m)	Nezer	0.25	0.12	2.95 – 7.81
	Tagon (Juin)	1.40	1.12	0.77 – 10.69
	Tagon (Août)	1.30	1.10	0.77 – 10.69
	Tagon (Février)	1.14	1.10	0.77 – 10.69
Sp (m)	Nezer	0.37	0.14	3.04 – 7.81
	Tagon (Juin)	0.98	0.92	1.31 – 8.77
	Tagon (Août)	0.89	0.83	1.31 – 8.77
	Tagon (Février)	1.03	0.84	1.31 – 8.77
Ht (m)	Nezer	1.5	0.5	9.3 – 24.7
	Tagon (Juin)	4.5	3.6	1.7 – 26.2
	Tagon (Août)	4.4	3.1	1.7 – 26.2
	Tagon (Février)	3.0	2.5	1.7 – 26.2
Nah (tree/ha)	Nezer	37	22	189 – 1253
	Tagon (Juin)	682	640	150 – 6729
	Tagon (Août)	677	586	150 – 6729
	Tagon (Février)	690	657	150 – 6729
Dbh(m)	Nezer	0.029	0.009	0.15 – 0.46
	Tagon (Juin)	0.086	0.062	0.02 – 0.56
	Tagon (Août)	0.081	0.065	0.02 – 0.56
	Tagon (Février)	0.064	0.058	0.02 – 0.56

Pour le site de Nezer, la LOO RMSE a été fortement améliorée, le gain est de 52% pour *Cd*, de 66% pour *Ht* et de 62% pour *Sp*. La précision était déjà élevée par régression simple pour Nezer avec un nombre limité de placettes ( $n = 12$ ) et la structure de la forêt est plus homogène dans des peuplements échantillonnés avec une gamme de variations réduite (âge > 12 ans). Ces conditions sont optimales pour produire un modèle linéaire simple, l'utilisation de modèles multiples dans ce cas là est discutable.

Pour le site du Tagon qui a une plus grande plage de variations dans les variables forestières et une plus grande variabilité de structure dans et entre les peuplements échantillonnés, l'amélioration est moins importante mais elle est tout de même significative. Pour l'image d'août par exemple, la LOO RMSE a été améliorée de 15,3 % pour *Cd*, de 29,5% pour *Ht* et de 6,7% pour *Sp*. Le gain en précision est meilleur pour *Dbh* ( 28%, 20%, 10% pour Tagon en Juin, Août et Février, respectivement) que pour *Cd* (20%, 15%, 3%). Même si *Cd* et *Dbh* sont fortement corrélés (*cf.* Figure 3.5(a)) *Cd* influence directement la texture de l'image contrairement au *Dbh*. Ce qui pourrait expliquer cette différence est une incertitude relative de mesure in situ supérieure pour *Cd* (voir les écarts-types dans les figures).

Le gain le plus élevé est pour la variable *Ht* (67%, 36%, 30%, 17% pour Nezer, Tagon en Juin, Août et Février, respectivement), probablement à cause de la plus grande complexité de la contribution de cette variable à la texture qui est liée à la longueur et la direction des couronnes et des ombres correspondantes telles que vues par le capteur. Le gain est le plus bas pour *Sp* sur le Tagon en été (6% en Juin et 7% en Août) mais il reste important pour l'image d'hiver (20% en Février).

Enfin, on peut remarquer que les résultats obtenus avec l'image d'hiver sont globalement meilleurs. Comme déjà dit précédemment lors de l'étude du modèle simple, on peut suspecter un effet de la faible hauteur du soleil (grandes ombres portées au sol) et du stade phénologique de la végétation du sous bois en hiver (surface foliaire verte minimale).

Enfin, les meilleures LOO RMSE observées pour le Tagon sont très satisfaisantes pour *Cd* et *Sp*, car elles sont très proches de la précision attendue par rapport à la résolution de l'image : LOO RMSE  $\sim 1,1m$  pour *Cd* et  $\sim 0,8m$  pour *Sp*, soit à peu près deux fois la résolution de Pan (0,5 m). Ainsi, pour les plus jeunes peuplements, dont la taille de l'arbre ou de l'espacement est inférieur à 1m, il est difficile de trouver un modèle offrant une bonne précision. Quelque soit la variable forestière forestière la meilleure LOO RMSE est de l'ordre de 10%

de l'étendue de la variation.

Ces résultats montrent que les solutions multiples et la méthode proposée de sélection de variables Random Prif sont adaptées pour produire des estimations robustes des variables forestières. Pour chaque site et chacune des dates, la méthodologie développée a permis de trouver des ensembles d'attributs de texture pertinents et adaptés à chacun des cas tout en restant robuste aux différentes conditions d'acquisition d'images, aux stades phénologiques, etc.

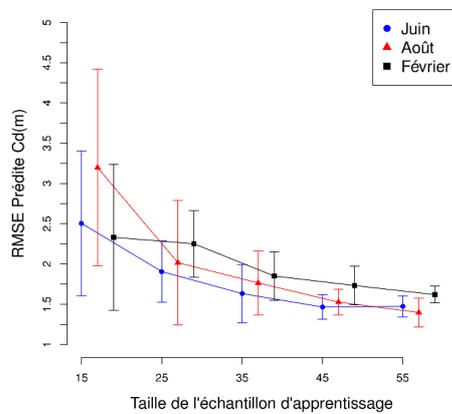
#### 4.4.3.4 Evaluation complémentaire de la précision des estimations

Cette dernière section est consacrée à la validation de notre méthode d'estimation des variables de structure forestière en terme de précision. Cette évaluation a été faite sur le site du Tagon où un grand nombre de mesures de terrain étaient disponibles ( $n = 111$ ). Ces mesures sont supposées être représentatives de l'ensemble de la variabilité des structures forestières de pin maritime sur la zone d'étude.

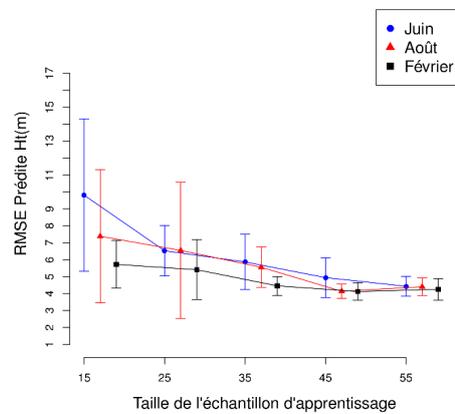
Les échantillons de terrain ont été divisés en deux ensembles de données indépendants: le premier ensemble est utilisé pour la sélection d'attributs de texture par Random PRiF (sur l'ensemble de données MS-Pan) et l'estimation des paramètres du modèle, le second ensemble est utilisé pour estimer l'erreur de prédiction (RMSE). La robustesse des prédictions est évaluée en répétant l'étape de modélisation sur différents ensembles d'apprentissage de taille croissante. Les deux sous-ensembles d'échantillons de terrain sont de même distribution (les échantillons ont été classés par rapport à leur diamètre de couronne et ont été répartis alternativement soit pour l'apprentissage soit pour le test). Le premier sous-ensemble contient 71 échantillons à partir desquels tous les échantillons servant pour l'apprentissage sont sélectionnés au hasard avec remise. Le second sous-ensemble contient les 40 échantillons restants est fixe et utilisé pour le test. L'erreur quadratique moyenne a été calculée sur les prédictions calculées sur l'ensemble test tout en augmentant la taille de l'ensemble d'apprentissage de 15 à 55 échantillons. L'erreur quadratique moyenne a été moyennée sur 10 répétitions pour chaque taille d'échantillon. Pour assurer une comparaison entre les trois dates, le même échantillonnage des données d'apprentissage et de test a été utilisé.

La figure 4.14 montre les résultats obtenus pour les trois dates sur les trois

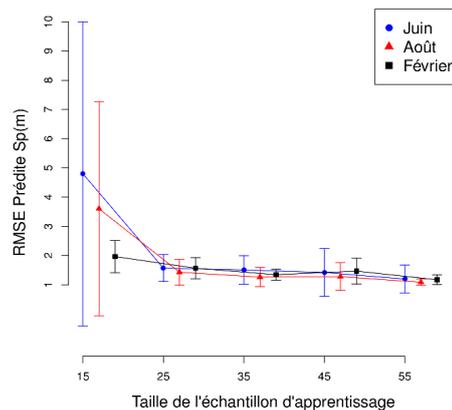
variables principales de la structure forestière:  $Cd$ ,  $Ht$  et  $Sp$ . L'observation de ces courbes montre que l'erreur de prédiction diminue de manière significative avec l'augmentation du nombre d'échantillons d'apprentissage, Pour l'ensemble des variables de structure forestière testées, on peut observer que la RMSE moyenne diminue jusqu'à atteindre un palier autour de 45 échantillons d'apprentissage. Ceci est une bonne indication quant au nombre minimum d'échantillons à considérer pour s'assurer une bonne performance de prédiction des principales variables forestières.



(a) Diamètre des couronnes ( $Cd$ )



(b) Hauteur des arbres ( $Ht$ )



(c) Espacement entre les arbres ( $Sp$ )

Figure 4.14 : Variation de la RMSE moyenne en fonction du nombre d'échantillons d'apprentissage. Résultats pour  $Cd$ ,  $Ht$  et  $Sp$  avec le même échantillon de test ( $n=40$ ) et 10 répétitions par taille d'échantillon d'apprentissage. Les barres représentent  $\pm 1$  écart type. Site du Tagon, images PLéiades.

Les écart-types diminuent aussi avec le nombre d'échantillons d'apprentissage mais ils peuvent rester importants. Ces résultats soulignent l'impact du choix des échantillons et de leur distribution sur les performances de prédiction. Aussi ils permettent de relativiser la qualité des résultats obtenus jusque là, cette expérience nous montre que l'on peut espérer estimer sur un secteur du massif forestier landais présentant une variabilité de la structure des peuplements similaire à celle du site du tagon. Ainsi, à partir d'une quarantaine d'échantillons mesurés sur le terrain, le diamètre des couronnes peut être estimé avec une erreur de prédiction moyenne de l'ordre de 1,7m, la hauteur des arbres avec une précision de l'ordre de 5m et l'espacement entre les arbres avec une précision de l'ordre de 1,5m.

#### 4.4.3.5 Généralisation à des études multi-annuelles

Des essais ont été faits pour essayer de prédire les variables de structure forestière à une date donnée, en utilisant le modèle de régression multiple calculé sur une autre date. Tous ces essais ont conduit à de mauvais résultats, ceci renforce notre idée selon laquelle l'optimisation des paramètres des attributs de texture est à faire pour chaque nouvelle image et nécessite donc des mesures de terrain associées. De nombreux paramètres non contrôlés (géométrie d'acquisition, saison...) influencent la texture de l'image et il ne semble pas possible de pouvoir généraliser notre approche pour des études multi-dates (calage d'un modèle sur une date et application du modèle à une autre date). Il serait par contre possible d'envisager, à terme, la constitution d'un catalogue canonique regroupant de nombreux échantillons d'images correspondant à différentes structures forestières, dans un grand nombre de situations différentes (angles de visée, angles solaires, saisons). Il serait alors possible, pour une image donnée, de trouver dans ce catalogue une situation semblable afin d'obtenir les bons attributs de texture et leurs paramétrages, ainsi que les coefficients associés.

## 4.5 Conclusion

Ce chapitre a permis d'appréhender la question de l'estimation des variables de structure forestière à partir de la texture des images à Très Haute Résolution spatiale. Parmi de nombreuses descriptions possibles de la texture d'image, l'étude s'est concentrée sur l'utilisation des matrices de co-occurrence (GLCM).

Les variables de structure forestière sont estimées à partir d'un attribut de

texture ou d'un ensemble d'attributs de texture complémentaires via régression linéaire simple ou multiple. Le principal apport de ce travail consiste à optimiser de manière automatique le paramétrage spatial des attributs de texture d'Haralick. Un grand nombre d'attributs de texture est calculé sur l'ensemble des données images disponibles (panchromatique et multispectrale, soit deux résolutions spatiales et différentes bandes spectrales) avec une vaste gamme de paramétrages possibles. C'est ensuite par sélection de variables que l'optimisation est faite, cela pour chaque variable de la structure forestière. L'étude a montré l'apport des régressions linéaires multiples, notamment sur le site du Tagon, ces dernières permettent de combiner le pouvoir explicatif des différents attributs de texture utilisés. Le problème de la colinéarité a été traité avec succès en proposant une nouvelle méthode de sélection de variables pour les régressions linéaires multiples. Cette nouvelle méthode, Random PRiF, permet de sélectionner, pour chaque variable forestière, un ensemble d'attributs de texture (issus des bandes multispectrales et de la bande panchromatique) performant, équilibré et plutôt dépourvu de colinéarité.

Les meilleures combinaisons d'attributs de texture combinent des attributs de texture issus des deux résolutions et avec différents paramétrages spatiaux. Ceci montre l'intérêt d'une approche multi-résolution et multi-échelle pour exploiter au mieux la texture des images THR permettant ainsi utilisation optimale des données de type Pléiades. Cette méthode a été appliquée avec succès sur les deux sites d'études, avec différentes configurations (angles de visée, angles solaires, saison). Les évaluations des performances confirment la robustesse de la méthodologie. Les performances de prédiction des différentes variables forestières étudiées sont très bonnes.

Bien entendu, l'utilisation de régressions linéaires est discutable. Des approches comme les régressions basées sur les arbres de décisions (tree-based regression), ou basées sur des approches non linéaires pourraient être envisagées. L'idée d'automatisation du processus de sélection des attributs et donc des meilleurs paramétrages est générique et pourrait être appliquée à d'autres attributs de texture, d'autres types de forêts ou bien à d'autres thématiques que celles sur la forêt. Enfin cette étude nous a permis de démontrer les capacités des images THR et notamment les images Pléiades pour la caractériser finement la structure des peuplements forestiers.

# Chapitre 5

## Cartographie des types de structures forestières

*Ce chapitre se focalise sur la mise en place d'une méthodologie opérationnelle pour la cartographie des types de structure forestière. Les résultats sont présentés sur les deux sites d'études en utilisant différents jeux de données THR disponibles. Une partie de ces résultats a fait l'objet de communications et publications (Beguet et al., 2013, 2014).*

### 5.1 Introduction

Ce chapitre présente notre réflexion sur le développement d'une stratégie opérationnelle pour la cartographie des types de structures forestières. La très haute résolution spatiale des images Pléiades doit permettre une cartographie fine des peuplements forestiers, en accédant à la variabilité intra parcellaire. Aussi, nous proposons de travailler au niveau pixel (et non au niveau de la parcelle, ou objet) afin d'accéder à une information spatialisée la plus fine possible. La classification des données de télédétection a longtemps fait l'objet de nombreuses études. En effet, les résultats de classification sont à la base de nombreuses applications environnementales et socio-économiques Tso and Mather (2001). Les scientifiques et les praticiens se sont beaucoup investis dans le développement d'approches de classification avancées, améliorant la précision et la rapidité des processus de classification Guo (2008, 2011); Lu and Weng (2007). En télédétection, de nombreuses caractéristiques peuvent être extraites des bandes spectrales originales, constituant souvent une information massive à gérer, en particulier si des caractéristiques de texture sont impliquées. La tâche de classification induit alors une grande complexité de calcul.

La sélection d'attributs peut être utilisée comme étape de pré-traitement pour la classification afin de fournir les attributs les plus pertinents en entrée des classifieurs. Elle réduit non seulement la complexité de calcul mais peut aussi augmenter la précision de classification, notamment pour les classes minoritaires Dash and Liu (1997); Guyon and Elissee (2003). En télédétection, la sélection d'attributs avant classification est généralement utilisée dans le contexte des données de grande dimension comme par exemple les images hyper-spectrales ou multi-sources Li et al. (2011); Serpico and Bruzzone (2001) ou encore, comme dans notre cas, l'analyse de texture Puig et al. (2010); Van Coillie et al. (2007). Cette notion de sélection de variables pour la classification a fait l'objet d'une attention particulière, notamment, nous proposons de tester le potentiel de notre méthode de sélection de variables basée sur la modélisation statistique des variables forestières pour la classification. Cette méthode sera comparée à une méthode de référence basée sur l'importance de variables des forêts aléatoires.

Une stratégie hiérarchique est proposée pour la classification des types de structure forestière, l'objectif étant de discriminer dans un premier temps les surfaces boisées (peuplement  $> 4$  ans) et les surfaces non boisées (coupes rases, régénérations récentes) avant de recalculer la texture uniquement pour les surfaces boisées pour ensuite les classifier. La pertinence de cette de cette approche, en comparaison avec une stratégie de classification directe, sera démontrée.

L'utilisation des forêts aléatoires, un classifieur ensembliste, nous permettra d'exploiter le concept de marge d'ensemble et d'en proposer une possible utilisation opérationnelle : la correction des erreurs d'étiquetage dans la base de donnée de référence.

Enfin, la robustesse de la méthodologie proposée sera évaluée sur le site du Tagon, en utilisant les 3 images Pléiades disponibles, à différentes périodes de l'année et avec différentes géométries d'acquisition.

## 5.2 Méthodes de sélection de variables pour la classification

### 5.2.1 Introduction

Les algorithmes d'apprentissage actif et de fouille de données doivent aujourd'hui traiter des problèmes de plus en plus complexes, où la quantité de données devient de plus en plus importante. Se concentrer sur les variables (ou attributs ou features) les plus pertinentes est donc devenu une nécessité pour de nombreux domaines d'application comme l'imagerie hyperspectrale ou l'analyse de texture d'images. La sélection de variables (feature selection) pertinentes s'oppose aux méthodes de d'extraction de variables (feature extraction) qui construisent des variables dérivées des variables initiales, comme par exemple le calcul d'indices, de ratios de bandes ou de combinaisons linéaires de variables comme pour les orthogonalisation type ACP (Analyse en composantes principales).

Les méthodes de sélection de variables permettent essentiellement de :

- améliorer les performances de prédiction.
- fournir des constructions de modèles plus rapides et praticables en terme de complexité.
- éviter l'over-fitting.
- mieux appréhender le processus de construction des données.

Les sous-ensembles ou le sous-ensemble optimal de variables peut toutefois contenir des variables corrélées entre elles et des variables à priori inutiles, Guyon and Elissee (2003) rappelle :

- La réduction du bruit et donc une meilleure séparation des classes peut être obtenue en ajoutant des variables présumées redondantes.
- Une très forte corrélation entre variables ne signifie pas une absence de complémentarité entre ces deux variables.
- Une variable qui est inutile seule peut améliorer les performances lorsqu'elle est en présence d'autres variables.
- Deux variables inutiles seules peuvent être très utiles lorsqu'elles sont combinées.

Idéalement, une méthode de sélection de variables devrait évaluer tous les sous ensembles de variables (ou attributs ou features)  $v_i$  possibles pouvant être générés à partir d'un ensemble de variables de départ  $D$  pour finalement conserver le meilleur sous ensemble de variable  $v_{best}$ . Cette procédure exhaustive est trop coûteuse et n'est pas praticable dès que la dimension de  $D$  devient grande. Différentes méthodes basées sur des heuristiques ou des recherches aléatoires tendent à réduire la complexité des calculs, cela en compromettant la performance de prédiction finale. Ces méthodes nécessitent un critère d'arrêt pour ne pas conduire une recherche exhaustive. Une méthode de sélection de variables typique nécessite quatre étapes (Dash and Liu, 1997) :

- une procédure de recherche, pour générer le sous-ensemble candidat suivant.
- une fonction objective ou fonction d'évaluation, pour examiner la qualité du sous-ensemble considéré
- un critère d'arrêt, pour décider l'arrêt du processus
- une procédure de validation, pour tester la validité du sous-ensemble considéré

Dash and Liu (1997) classe les procédures de recherche en trois catégories principales : recherche exhaustive, recherche basée sur des heuristiques et enfin recherche aléatoire. Les fonctions d'évaluation peuvent être des distances, des dépendances, des pertinences ou des mesures d'erreur de classification.

Dans le contexte de la classification, les techniques de sélection de variables peuvent être organisées en trois catégories. Ces catégories diffèrent dans la manière de combiner la sélection de variables avec la construction du modèle, on distingue les méthodes *filter*, les méthodes *wrapper* et les méthodes *embedded*.

Dans un premier temps nous détaillerons les caractéristiques de ces différentes méthodes, dans un second temps seront présentées les méthodes les plus couramment utilisées pour des problèmes de classification d'images et notamment d'images de télédétection.

### **5.2.2 Bases théoriques, taxonomie des méthodes de sélection de variables :**

Cette section présente les trois grandes familles de techniques de sélection de variables. Un tableau reprenant les méthodes les plus couramment rencontrées avec leurs avantages et inconvénients est ensuite proposé afin d'aider au choix

d'une méthode de sélection de variables en fonction de la complexité et de la disponibilité des données.

### 5.2.2.1 Méthodes *Filters* :

Ces méthodes sont une étape préliminaire à la classification, pour réduire la dimension de l'espace des variables et éviter l'over-fitting. Elles sont indépendantes du type de classifieur utilisé et sont très rapides (Guyon and Elissee, 2003). Elles sont détachées du processus d'apprentissage des classifieurs. Elles sont adaptées à des classifieurs non supervisés. Cependant, elles ignorent les interactions entre les variables.

Leur principe général est de classer les variables selon un critère de pertinence, autrement dit elles donnent un poids à chacune des variables en regardant uniquement les propriétés intrinsèques des données. Les variables avec un faible score sont alors enlevées de l'espace de départ, cet espace réduit est passé en entrée du classifieur. La sélection de variables est donc faite une seule fois et différents classifieurs peuvent alors être évalués. La plupart de ces méthodes sont univariées, cela signifie que chaque variable est considérée séparément des autres comme par exemple la méthode basée sur le F-test Mallet (2010) ou l'algorithme Relief Kira and Rendell (1992).

Ce problème a poussé à la création de méthodes filter multivariées qui introduisent à un certain degré les dépendances entre variables comme l'algorithme ReliefF Kononenko (1994); Robnik-Sikonja and Kononenko (2003). Les filtres à base de la théorie de l'information comme l'algorithme Markov blanket Koller and Sahami (1996); Zhang (2010) constituent une famille importante. Comme mentionné par Bruzzone and Serpico (2000), de nombreux critères ont été définis pour le problème de la sélection d'attributs adaptée aux cas multi-classes. Critères de divergences, mesures de distances et critères basés sur les «scatter matrices» sont les plus couramment utilisés.

**5.2.2.1.1 Relief et ReliefF :** L'idée de l'algorithme Relief est d'estimer les attributs selon la manière dont ses valeurs se distinguent parmi les instances qui sont proches l'une de l'autre. Pour cela, Relief cherche pour une instance donnée  $R$  ses deux instances les plus proches (Nearest Neighbours), l'une  $H$  qui appartient à sa classe et l'autre  $M$  qui appartient à une classe différente. L'algorithme

Relief choisit aléatoirement  $m$  instances à partir de l'ensemble d'apprentissage ( $m$  est défini par l'utilisateur). Relief utilise une fonction *diff* pour calculer la différence entre les valeurs d'un attribut dans les deux instances. Le poids  $W(A)$  est l'estimation de la qualité d'un attribut  $A$ . Différentes distances peuvent être envisagée pour l'utilisation de l'algorithme Relief.

L'algorithme ReliefF est une extension de Relief proposé par Kononenko (1994). Il étend l'algorithme original sur plusieurs points :

- il n'est pas limité au problème de deux classes
- il traite le problème des valeurs manquantes.

Le choix de l'instance la plus proche est d'importance cruciale dans Relief. L'objectif est de trouver les instances les plus proches en respectant les attributs pertinents. Pour augmenter la fiabilité de l'approximation de probabilité, ReliefF cherche les  $K$  instances les plus proches de l'instance à traiter de chaque classe au lieu d'une seule instance la plus proche de chaque classe.  $K$  est un paramètre défini par l'utilisateur ; l'expérimentation faite par Kononenko (1994) montre que quand  $K$  vaut 10, les résultats sont satisfaits empiriquement.

ReliefF met à jour les estimations de qualité des attributs de la manière suivante : Pour l'instance  $Ri$  , il cherche les  $k$  instances  $Hj$  les plus proches, au lieu d'une seule instance. Il prend également en compte la probabilité de chaque classe  $P(C)$ , estimée à partir de la base d'apprentissage, où  $C = class(Ri)$ . ReliefF a étendu la fonction  $diff(A, Instance1, Instance2)$  pour prendre en compte un attribut ayant des valeurs manquantes lors de l'estimation de sa qualité.

### 5.2.2.2 Méthodes *Wrappers* :

Les méthodes *wrappers* ont été popularisées par Kohavi and Jhon (1997). Ces méthodes utilisent l'algorithme de classification comme fonction d'évaluation. La minimisation de l'erreur de prédiction permet de trouver le meilleur sous ensemble de variables.

La première chose est donc de choisir le classifieur supervisé à utiliser avant de trouver l'espace de tous les sous-ensembles de variables à considérer. Il faut ensuite pouvoir accéder à la performance de prédiction de l'ensemble d'apprentissage pour guider la recherche et la stopper. Pour les problèmes de faible dimensionnalité, une recherche exhaustive est possible mais le problème est connu comme étant NP-complet et par conséquent l'exhaustif est impraticable dans la majorité

des cas.

Ces méthodes sont souvent critiquées car elles nécessitent une grande quantité de calculs mais des alternatives ont été proposées. Deux grandes familles distinctes existent : les approches déterministes et les approches aléatoires qui diffèrent par leur procédure de sélection. Pour trouver l'espace de tous les sous-ensembles à tester, cette procédure de recherche ou algorithme de recherche « emballe » (to wrap) le modèle de classification. L'un des inconvénients majeur de ces méthodes est qu'elles produisent un plus grand risque d'over-fitting que les méthodes filter, sans oublier le coût de la construction du classifieur (Sayes et al., 2007). Les approches les plus courantes sont les SFS (Sequential Forward Selection) et SBS (Sequential Backward Selection). D'autres solutions plus élaborées ont été proposées, notamment par Serpico and Bruzzone (2001). Il propose de nouvelles stratégies de recherche sous-optimale de variables, le « Steepest Ascent Algorithm » (SA) et le « Fast Algorithm for a Constraint Search » (FCS). Les résultats expérimentaux ont montré la performance de ces stratégies comparées aux classiques SFS et SFFS (Sequential Forward Floating Selection).

### 5.2.2.3 Méthodes *Embedded* :

Ces méthodes de sélection de variables sont ici intégrées dans le processus d'apprentissage et sont donc plus rapides que les méthodes wrappers. La recherche du sous-ensemble optimal est faite durant la construction du classifieur, c'est une recherche combinée dans l'espace des sous-ensembles possibles et des hypothèses. Elle permettent une meilleure utilisation des données car elles n'ont pas besoin de refaire un apprentissage pour chaque sous-ensemble de variables. Les arbres de décisions et les forêts aléatoires sont les plus répandus (Breiman, 1996, 2001; Genuer et al., 2010) ainsi que l'utilisation du vecteur de poids pour les Support Vector Machine (Guyon and Elissee, 2003; Li et al., 2011). Les forêts aléatoires seront détaillées dans la section 5.3.

Pour la sélection de variables basée sur l'importance de variable des forêts aléatoires, l'importance d'un attribut peut être estimée en permutant aléatoirement toutes les valeurs de cet attribut dans les échantillons OOB (Out Of Bag) pour chaque arbre de la forêt aléatoire. OOB se réfère aux échantillons qui n'ont pas été utilisés durant le *bootstrap* sur les échantillons d'apprentissage. Pour d'autres méthodes ensemblistes, l'échantillon OOB peut être remplacé par un échantillon de données de validation, différent de celui de l'apprentissage. La

sélection d'attributs par l'importance des forêts aléatoires imite l'absence d'un attribut grâce à sa permutations aléatoire. L'importance d'un attribut est alors déterminée par la perte de précision causée par son absence Breiman (2001). Cette mesure d'importance par permutation à déjà été utilisée avec succès pour la cartographie urbaine Guo (2011); Guo et al. (2011). Toutefois, la sélection d'attributs via la classification supervisée est non seulement coûteuse mais elle est aussi dépendante du classifieur sous-jacent.

### 5.2.3 Applications en classification d'images de télédétection

Cette section présente des exemples récents d'applications de méthodes de sélection de variables. La sélection de variables est surtout utilisée en cas de grande dimensionnalité. On la retrouve souvent dans la classification des images hyperspectrales (Guo, 2008; Li et al., 2011) ou bien dans les applications utilisant de nombreux attributs tels que la classification de texture (Puig et al., 2010; Van Coillie et al., 2007).

Pour réduire la dimensionnalité des données hyperspectrales, Li et al. (2011) propose une stratégie hybride basée sur les algorithmes génétiques et les SVM (GA-SVM). Cette méthode de type wrapper cherche la meilleure combinaison de bandes en terme de performance de classification. Les bandes sont d'abord groupées à l'aide d'un critère d'information mutuelle entre bandes adjacentes puis les groupes les moins pertinents sont supprimés afin de réduire le coût de la recherche par algorithme génétique. Toujours dans le cas d'utilisation de données hyperspectrales, Guo (2008) utilise l'information mutuelle dans une stratégie avec optimisation progressive (séquentielle). Cette méthode entre donc dans la catégories des méthodes filter. Serpico and Bruzzone (2001), toujours pour réduire la dimensionnalité de données hypersperctrals propose de nouvelles stratégies efficaces de recherche sous-optimale de variables (SA, FCS).

Afin de cartographier la forêt à l'aide de la texture d'images à Très Haute Résolution, Van Coillie et al. (2007) combine algorithmes génétiques et réseaux de neurones pour la classification, montrant le potentiel de cette méthode lorsque peu de données d'apprentissages sont disponibles. Kerroum et al. (2010) propose une méthode de sélection de variables basée sur l'information mutuelle pour la cartographie forestière à partir d'une image multispectrale Spot. Les attributs de texture extraits des transformées en ondelettes et des matrices de co-occurrence

sont ainsi sélectionnés en amont d'un classifieur SVM.

Ce rapide tour d'horizon est ici pour rappeler que si l'étude des techniques de méthodes de sélections de variables est ancienne, il n'y a pas aujourd'hui de méthode faisant l'unanimité. Par conséquent l'utilisateur se doit de sélectionner la méthode la plus adaptée à son problème et à ses capacités en terme de puissance de calcul. Il n'y a pas à priori de technique de sélection de variables optimale. De nombreux facteurs interviennent dans le choix d'une méthode de sélection de variables, notamment le nombre de variables de l'ensemble de départ, la disponibilité de données d'apprentissage, le classifieur envisagé ou encore la puissance de calcul disponible. Si les méthodes filtres sont simples, rapides, efficaces et indépendantes du classifieur, elles sont moins performantes que les méthodes wrapper. Ces dernières supposent que l'on a choisi le classifieur à utiliser, reste à choisir la méthode de sélection la moins coûteuse (si peu de puissance de calcul est disponible) ou la plus optimale (dans le cas d'une dimensionnalité raisonnable). Enfin les méthodes intégrées au processus d'apprentissage sont performantes et relativement légères, elles supposent cependant la disponibilité d'une bonne base de données d'apprentissage.

*If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems. R.L. Plackett (1979, discussion)*

## 5.3 Les forêts aléatoires : classifieur ensembliste

Les Random Forests ou forêts aléatoires ont été introduites par Breiman (2001). C'est un algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression.

L'algorithme appartient à la famille des méthodes d'ensemble de classifieurs (Dietterich, 2000). Le principe général de ces méthodes est de construire une collection de classifieurs (prédicteurs) faibles, pour ensuite agréger l'ensemble de leurs prédictions afin d'obtenir un classifieur robuste. Le classifieur faible considéré est un arbre de décision de type *Classification and Regression Tree* (CART) (Breiman et al., 1984) dont le principe général est de partitionner récursivement l'espace d'entrée de façon dyadique, puis de déterminer une sous-partition optimale pour la prédiction.

Les forêts aléatoires peuvent atteindre une précision de classification comparable au boosting (Breiman, 2001), ou aux SVM (Pal, 2005; Zhu, 2008). Ce classifieur est robuste à l'overfitting, efficace pour de grands volumes de données (imagerie hyperspectrale, grand nombre d'attributs,..) avec des temps de calculs rapides. Il ne nécessite pas d'hypothèses sur la distribution des données, ce qui est intéressant lorsque différents types de variables d'entrée sont utilisées. Par ailleurs, il permet de calculer une estimation interne non biaisée de l'erreur globale au fur et à mesure de la construction de la forêt (phase d'apprentissage) et permet également d'estimer l'importance de chaque variable pour l'application donnée.

Ces propriétés exceptionnelles le rendent approprié pour la classification de la télédétection. Il a été appliquée avec succès aux données multispectrales (Pal, 2005), aux images SAR multitemporelles (Waske and Braun, 2009), aux images hyperspectrales (Ham et al., 2005), ou encore aux données multi-sources (Gislason et al., 2006; Guo et al., 2011; Waske and Benediktsson, 2007).

### 5.3.1 Algorithme

Le classifieur des forêts aléatoires possède deux arguments principaux, notés ici  $n_{tree}$  et  $m_{try}$  qui correspondent respectivement au nombre d'arbres dans l'ensemble de classifieurs et le nombre de variables considérées à chaque nœud de l'arbre.

1. Tirer  $n_{tree}$  échantillons aléatoires avec remise parmi l'échantillon d'apprentissage (appelés bootstraps)
2. Pour chaque échantillon, construire un arbre de décision de type CART de la façon suivante : Pour chaque nœud, au lieu de choisir le meilleur découpage pour toutes les variables, on prend un tirage aléatoire de  $m_{try}$  variables pour former la décision associée au nœud. Le meilleur découpage est basé sur le calcul de la pureté du nœud (Gini impurity) (Breiman, 2001).
3. Prédire les nouvelles données en agrégeant les prédictions des  $n_{tree}$  arbres (En général, la majorité des votes dans le cas d'une classification, la moyenne pour la régression).

Il a été démontré que l'erreur d'estimation dépend de deux points essentiels:

- La corrélation entre deux arbres de la forêt: Augmenter la corrélation augmente l'erreur.
- La robustesse de chaque arbre de la forêt: Un arbre avec un faible taux d'erreur est un bon classifieur. Le fait d'augmenter la robustesse d'arbres individuels diminue le taux d'erreurs de la forêt.

Concernant le paramétrage du classifieur, réduire la valeur de  $m_{try}$  revient à diminuer la corrélation et la robustesse. Il faut donc trouver un intervalle optimal pour  $m_{try}$ . Dans la plupart des cas, un choix optimal sera  $\sqrt{p}$ , où  $p$  représente le nombre total de variables. Ce choix se base sur la minimisation de l'erreur Out-Of-Bag (OOB).

Cet algorithme est l'un des plus performants pour la classification de données, nous avons donc privilégié son utilisation dans cette étude. De plus, il a l'avantage de fournir une indication sur la fiabilité de la classification qui est la marge d'ensemble. Cette information peut avoir une grande utilité d'un point de vue opérationnel, nous présenterons par la suite une de ses possibles applications.

### 5.3.2 Erreur OOB

En plus de construire un prédicteur, l'algorithme des forêts aléatoires calcule une estimation de son erreur de généralisation : l'erreur OOB. "Out-Of-Bag" signifie ici "en dehors du bootstrap". Le procédé de calcul de cette erreur est le suivant:

Fixons une observation  $(x_i ; y_i)$  de l'échantillon d'apprentissage  $L_n$ . Considérons maintenant l'ensemble des arbres construits sur les échantillons bootstrap ne contenant pas cette observation, c'est-à-dire pour lesquels cette observation est "Out-Of-Bag" (environ un tiers des observations).

Nous agrégeons alors uniquement les prédictions de ces arbres pour fabriquer notre prédiction de  $Y_i$ . Après avoir fait cette opération pour toutes les données de  $L_n$ , nous calculons alors l'erreur commise par nos prédictions. Dans le cas d'une classification, la proportion d'observations mal classées est  $\frac{1}{n} * \sum_{i=1}^n 1_{\hat{y} \neq y_i}$ . Cette quantité est appelée erreur OOB du prédicteur des forêts aléatoires.

Cette estimation s'impose les mêmes contraintes que celles des estimateurs classiques de l'erreur de généralisation (échantillon test, validation croisée), au

sens où les données prédites sont des données qui n'ont pas été rencontrées au préalable par le prédicteur utilisé.

Un avantage de l'erreur OOB par rapport aux estimateurs classiques est qu'elle ne nécessite pas de découpage de l'échantillon d'apprentissage. Ce découpage est en quelque sorte inclus dans la génération des différents échantillons bootstrap. Cependant, il faut bien noter que pour chaque observation ce n'est pas le même ensemble d'arbres qui est agrégé. En conséquence, cette erreur estime l'erreur de généralisation d'une forêt, mais elle n'utilise jamais les prédictions de la forêt elle-même, mais plutôt celles de prédicteurs qui sont des agrégations d'arbres de cette forêt.

Un inconvénient de l'erreur OOB est qu'elle est souvent considérée comme étant trop optimiste. Néanmoins, dans ce rapport, nous utilisons uniquement l'erreur OOB pour comparer des prédicteurs entre eux, et non pour obtenir une estimation précise de leurs erreurs de généralisation.

### 5.3.3 Marge d'ensemble

Le concept de marge d'ensemble a été introduit par Schapire et al. (1997); Schapire and Singer (1998) pour expliquer le succès des algorithmes de type *boosting*. Le concept a ensuite été généralisé pour les autres types de classifieurs ensemblistes (Breiman, 2001). Se basant sur le principe que les échantillons d'apprentissage sont des paires de la forme  $(x_i, y_i)$  où  $x_i$  est une instance et  $y_i$  son étiquette de référence. La marge  $m_i$  de l'instance  $x_i$  est calculée de la manière suivante :

$$m_i = \text{margin}(x_i, y_i) = \frac{v_{i,y_i} - \sum_{c \neq y_i} v_{i,c}}{\sum_c v_{i,c}} \quad (5.1)$$

où  $v_{i,y_j}$  est le nombre de votes pour la classe "vraie"  $y_j$  et  $v_{i,c}$  est le nombre de votes pour toutes les classes  $c$  avec  $c \neq y_i$ . De plus; la marge est donnée par la différence entre le nombre des votes corrects et des votes incorrects. La marge mesure la force du vote, elle varie entre -1 et +1. Pour un échantillon, une marge positive indique qu'il a été correctement classifié tandis qu'une marge négative indique qu'il n'a pas été correctement classifié. Plus la marge est grande (en valeur absolue) plus la fiabilité, ou confiance, de la classification est grande.

La marge d'un échantillon révèle certaines de ces caractéristiques. Une forte valeur positive signifie que la majorité des arbres ont bien classé cet échantillon, ce qui indique qu'il se situe au centre de la distribution de la classe concernée.

Au contraire, une forte valeur négative indique que peu d'arbres ont bien classifié l'échantillon ce qui indique que cet échantillon peut représenter du bruit ou être un outlier relativement à cette classe. Une valeur de marge proche de 0 indique une faible fiabilité, ou confiance, de la classification. Ceci indique que cet échantillon est probablement à la frontière entre deux classes. De nombreuses études ont montré que la généralisation de la performance d'un classifieur ensembliste est lié à la distribution de ses marges sur les échantillons d'apprentissage. Dans l'étude qui suit, la marge sera exploitée comme une mesure de confiance du classifieur Random Forest. Elle sera utilisée pour une possible application opérationnelle à savoir la correction des erreurs d'étiquetage. En effet, considérer un ensemble de données étiquetées comme une "vérité" est une erreur, qui plus est lorsque, comme dans notre cas, cet étiquetage est basé sur une photo-interprétation humaine. On verra que l'observation et l'interprétation de cette information de marge d'ensemble nous permettra de détecter des erreurs d'étiquetage dans la base de donnée de référence.

## 5.4 Méthodologie : vers une stratégie opérationnelle

L'objectif principal de ce chapitre est de développer une méthodologie pour la classification des types de structure forestière, tournée vers l'opérationnel, et d'en évaluer les performances à l'aide des trois images sur le site du Tagon. Comme vu dans le chapitre précédent, l'idée principale est d'automatiser le paramétrage des attributs de texture issus des GLCM via une méthode de sélection de variables appropriée en amont du classifieur, basée sur de petits échantillons d'images. Les meilleurs attributs de texture seront ensuite calculés sur l'image entière (uniquement à l'intérieur des parcelles fournies sous forme de SIG) avant d'être classifiés par les forêts aléatoires pour réaliser une cartographie des types de structures forestières. Le schéma présenté figure 5.1 résume notre stratégie opérationnelle.

Dans un premier temps, les attributs de texture sont calculés sur les images correspondantes aux placettes de mesures observées in-situ. Comme dans le chapitre précédent, ces attributs de texture sont calculés aux deux résolutions (Panchromatique et Multispectrale) en vue d'être combinés par la suite. Notre stratégie est hiérarchique : elle propose dans un premier temps de sélectionner, de manière spécifique, par la méthode de l'importance de variable des forêts aléatoires, les meilleurs attributs pour discriminer la classe 1 (forêt en mutation) des

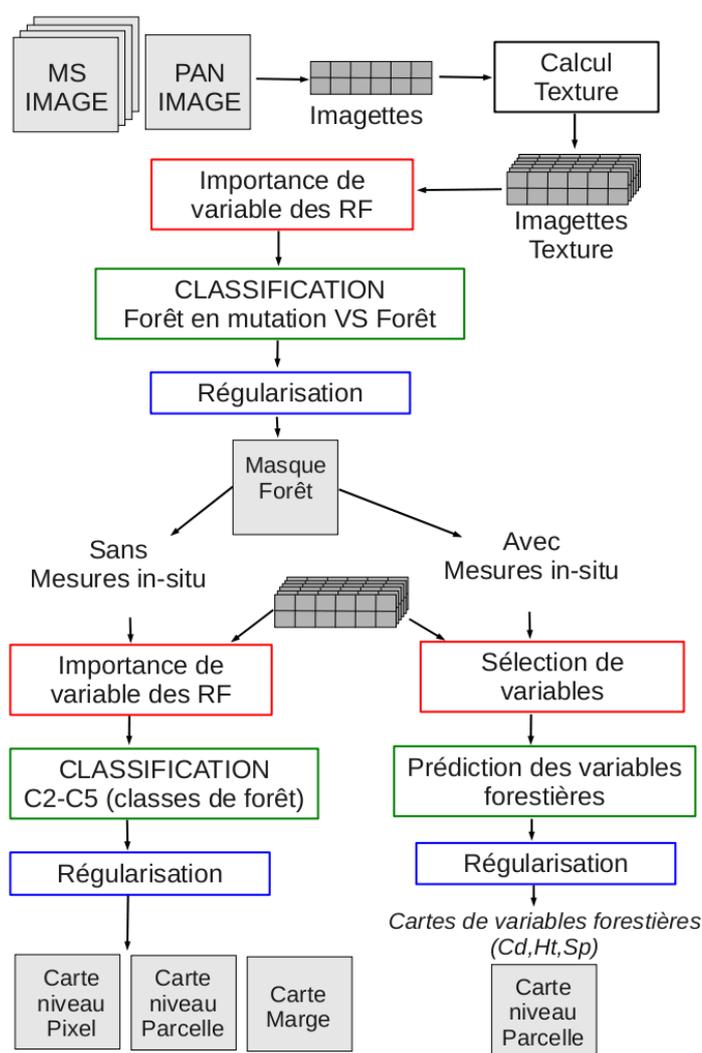


Figure 5.1 : Représentation schématique de la stratégie opérationnelle globale envisagée.

autres classes (forêt). Après régularisation, un masque de forêt est obtenu.

Une deuxième passe de sélection d'attributs est alors effectuée, spécifique aux pixels de forêt, les meilleurs attributs de texture obtenus sont ensuite calculés avant d'être classifiés. Comme l'illustre la section 5.6.1, calculer la texture sans prendre en compte les pixels qui ne sont pas de la forêt permet d'optimiser la finesse des attributs de texture et donc d'améliorer la reconnaissance des types de structures forestières.

La figure 5.1 rappelle que dans le cas où des mesures dendrométriques sont disponibles, toute la méthodologie vue au chapitre précédent peut s'appliquer afin de fournir des cartes de variables dendrométriques à l'échelle du peuplement. Dans ce chapitre, nous nous sommes concentrés sur le potentiel d'une classification au niveau pixel, ceci afin d'exploiter au mieux la résolution fine des images Pléiades et de distinguer des variations intra-parcellaires. Le potentiel des attributs de texture sélectionnés sur la base de la modélisation statistique des variables forestières pour réaliser la classification d'image est testé dans la section 5.5.

Nous avons choisi de produire des cartes à la résolution de 2m (résolution Multispectrale), ceci étant suffisant d'un point de vue thématique. De plus travailler à cette résolution permet de réduire considérablement le volume de données à traiter, facilitant ainsi la production de nombreux tests. Pour tous les tests combinant attributs de texture panchromatique et multispectraux, la texture panchromatique est bien entendu calculée à pleine résolution (0,5m) avant d'être ré-échantillonnée à la résolution de 2m. Ceci permettant de combiner les deux résolutions et ainsi optimiser l'utilisation de la donnée et du potentiel multi-résolution et multi-échelle.

Pour tous les tests (exceptées les expériences sur le site de Nezer, détaillées plus loin), les forêts aléatoires ont été utilisées avec 100 arbres de décision (aussi bien pour la sélection d'attributs que pour la classification). L'influence de ce paramètre  $n$  a pas été testée dans cette étude. Les échantillons utilisés pour la sélection d'attributs appartiennent à la base de référence correspondant aux mesures faites in situ. Pour l'apprentissage et le test des classifications, les 200 placettes étiquetées par photo-interprétation sont utilisées, 100 placettes pour l'apprentissage et 100 placettes pour le test (réparties de manière aléatoire, les deux ensembles ont donc la même distribution).

Pour l'étape de régularisation, nous avons choisi de régulariser l'image de classification par un vote majoritaire dans un voisinage de taille fixe. Le rayon de ce voisinage, pour toute la suite de cette étude, a été fixé à 4 pixels MS, cette dimension correspond à la taille minimale d'une trouée, c'est à dire supérieure à l'espacement maximum normal entre les arbres. Enfin, chaque étape de l'approche hiérarchique sera évaluée séparément (C1 Vs. Autres puis C2 à C5), cela pour les trois dates disponibles sur le site du Tagon. Ces résultats seront ensuite fusionnés pour obtenir une estimation de la précision globale de classification que l'on peut

attendre de notre méthodologie de cartographie des types de structure forestières à partir d'images Pléiades.

## 5.5 Recherche d'une méthode de sélection d'attributs optimale

Cette section a pour objectif de présenter le déroulement de notre réflexion sur la sélection des attributs de texture à passer en entrée du classifieur. Une méthode de sélection d'attributs de texture pour l'estimation des variables de la structure forestière ayant été développée dans le chapitre précédent, nous avons voulu tester le potentiel de cette sélection d'attributs basée sur la modélisation et donc l'apport de variables biophysiques pour la construction de bons classifieurs. D'autre part, nous avons choisi la méthode de sélection d'attributs basée sur l'importance de variables des forêts aléatoires comme méthode de référence. Comme dans le chapitre précédent, les premiers tests méthodologiques ont été réalisés sur le site de Nezer, dans l'attente des images Pléiades.

### 5.5.1 Expérience sur le site de Nezer

Cette première expérience a été menée sur l'image Quickbird multispectrale (acquise en 2003). Comme dans la section 4.4.1, 6 classes de structure forestière sont définies, dont 5 classes de forêt numérotées ici de C1 à C5. Ces premiers tests ont été réalisés en calculant la texture sur l'image d'intensité I. Les attributs de textures et les paramétrages associés sont identiques à ceux présentés section 4.3.2. Ces attributs sont calculés sur les mires présentées section 4.3.3. Ces mires sont utilisées pour les deux familles de méthodes de sélections d'attributs, celles basées sur la modélisation des variables de la structure forestières (LARS stepwise et random PRiF seront ici testées) et la méthode de référence utilisant l'importance de variables des forêts aléatoires. Seules les variables  $Cd$ ,  $Ht$  et  $Sp$  sont considérées dans cette expérience.

Dans ce qui suit, nous montrerons le potentiel de notre approche basée sur la modélisation des variables forestières pour la cartographie des types de structures forestière en utilisant deux classifieurs supervisés bien connus : les *Bagging Trees* (BT), une méthode d'ensemble d'arbres de décision, comme les RF (*Random Forest*) ou forêts aléatoires Breiman (1996) (pour être cohérent avec la méthode de sélection de variable basée sur les RF) ainsi que les machines à vecteurs de sup-

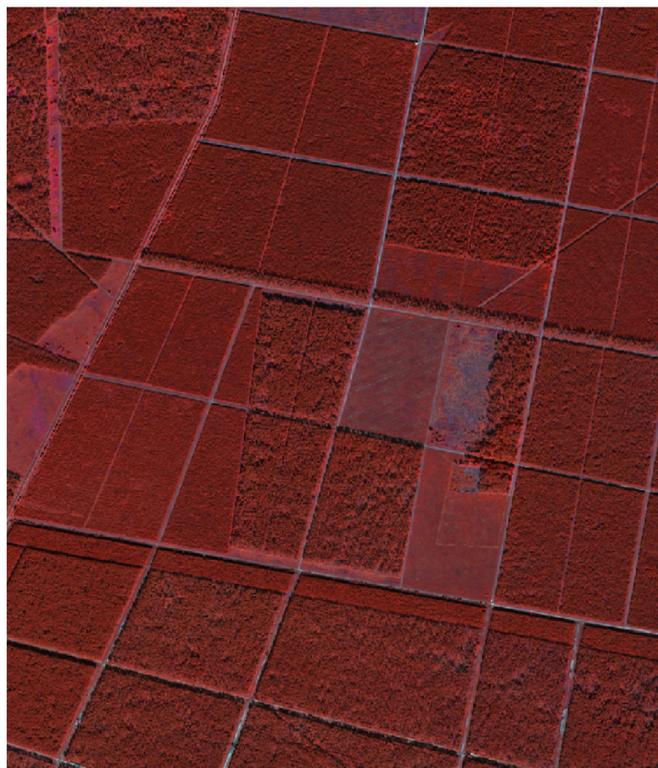


Figure 5.2 : Image multispectrale Quickbird (ROI 2),  $1136 \times 972$  pixels, forêt de Nezer (composition colorée PIR, R, V).

port (Support Vector Machines) ou SVM, un autre classifieur puissant, largement utilisé en télédétection Mountrakis et al. (2011).

Deux ROIs (Region Of Interest) distinctes de taille respectives  $1109 \times 1050$  pixels et  $1136 \times 972$  pixels (figure 5.2), sont extraites de l'image Quickbird et testées indépendamment. Seuls les pixels de forêt sont considérés, le masque utilisé et les peuplements de référence ont été fournis sous forme de SIG. Bien sûr, ces régions d'intérêt sont indépendantes de la mire impliquée dans les processus de sélection d'attributs. Chacun des deux jeux de données a été divisé en deux parties égales, un échantillon d'apprentissage et un échantillon de test, pour la classification supervisée.

#### 5.5.1.1 Résultats

Notre méthode de sélection d'attributs par modélisation est comparée à la méthode de sélection par importance de variables des forêts aléatoires (RF pour Random Forest) utilisée avec 100 arbres de décision. Les deux méthodes de sélection

d'attributs sont appliquées sur la même mire. Notre approche automatique par modélisation nous a amené à considérer des sous-ensembles de 3 attributs pour chaque variable forestière (considérant la statistique  $C_p$  de Mallow Efron et al. (2004)), nous retenons donc un sous-ensemble composé de 3 attributs pour la méthode de sélection d'attributs des RF. Un ensemble de 3 attributs de texture composé du meilleur attribut de texture pour chacune des variables forestières est aussi testé (et nommé  $CDHTSP$ ). Ainsi, un descripteur de 7 attributs est considéré pour la classification pixel: les quatre bandes spectrales (R, G, B, PIR) et trois caractéristiques de texture fournies par l'étape de sélection d'attributs. La classification est effectuée par les Bagging Trees (BT) avec un nombre d'arbres fixé à 100 et les SVM, avec un noyau linéaire.

Ces résultats présentés tableau 5.1 montrent le potentiel des attributs sélectionnés par modélisation statistique pour la classification des types de structure forestière. En effet, les deux méthodes de sélection de variables permettent d'obtenir des performances similaires. La méthode de sélection par modélisation statistique Random PriF est plus performante que la méthode LARS stepwise dans la plupart des cas. Les sélections basées sur l'estimation de la variable forestière  $Sp$  (espacement entre les arbres) fournissent de meilleures performances que les sélections basées sur les deux autres variables forestières ( $Cd$  et  $Ht$ ). Enfin, la combinaison impliquant le meilleur attribut pour chacune des variables forestières apparaît être la stratégie la plus performante. Ces considérations sont toutefois à relativiser étant données les très faibles différences en termes de précisions globales.

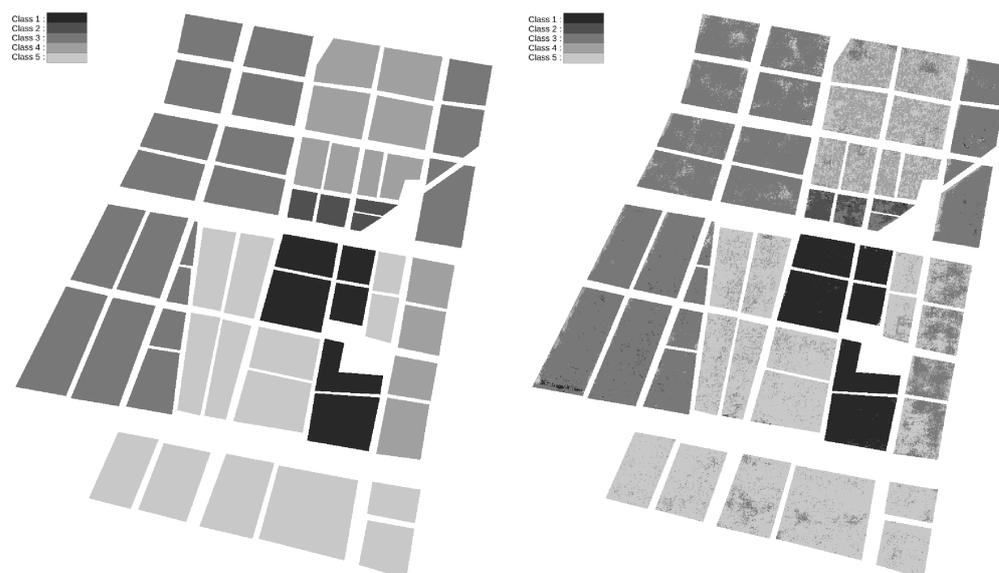
Tableau 5.1 : Performances de classification globale (%) par différentes stratégies de sélection de variables sur deux échantillons de l'image multispectrale Quickbird, avec 3 attributs de texture.

	Roi1		Roi2	
	BT	SVM	BT	SVM
<b>RF importance</b>	<b>74.71</b>	<b>75.52</b>	<b>82.43</b>	<b>83.53</b>
LARS stepwise Cd	70.52	70.44	77.43	78.03
Random PRiF Cd	71.98	70.78	78.79	77.88
LARS stepwise Ht	75.32	<b>75.71</b>	81.95	82.17
Random PRiF Ht	<b>77.52</b>	74.75	82.4	80.56
LARS stepwise Sp	76.54	76.47	83.35	83.1
Random PRiF Sp	<b>76.8</b>	<b>77.64</b>	83.98	<b>83.91</b>
LARS stepwise CDHTSP	75.51	76.21	83.47	83.23
Random PRiF CDHTSP	<b>77.97</b>	<b>77.98</b>	<b>83.97</b>	<b>83.84</b>

Le tableau 5.2 présente la matrice de confusion obtenue pour la classification de la ROI2 par Bagging Trees en utilisant les trois meilleurs attributs de texture pour la variable Sp. Elle correspond à la classification présentée figure 5.3. Cette matrice de confusion nous informe sur les difficultés rencontrées par le classifieur. On peut noter que deux classes sont moins bien reconnues (C2 et C4). La mauvaise reconnaissance de C2 peut être due à un trop faible nombre d'observations. On peut alors imaginer qu'un dispositif expérimental plus équilibré permettrait d'améliorer la reconnaissance de cette classe. Par ailleurs, la frontière entre la classe C4 et la classe C5 est très difficile à déterminer, ce qui explique la grande confusion commise pour la classe C4.

Tableau 5.2 : Matrice de confusion pour la classification de l'échantillon ROI 2 par Bagging Trees utilisant les 3 meilleurs attributs de texture décrivant la variable Sp (espacement entre les arbres)

	C1	C2	C3	C4	C5	Total	%Erreur
C1	29468	310	62	6	40	29886	1.39
C2	114	2506	1981	7	2	4610	45.64
C3	275	504	101323	4398	1075	107575	5.81
C4	7	17	5131	23406	18366	46927	50.12
C5	14	7	1022	10204	74460	85707	13.12
						Taux d'erreur total	23.21



(a)Echantillons de référence (5 classes de structure forestière)

(b)Classification Bagging Trees utilisant les 3 meilleurs attributs de texture décrivant la variable Sp (espacement entre les arbres)

Figure 5.3 : Classification de l'image multispectrale Quickbird (ROI 2) du site de Nezer utilisant la sélection d'attributs par modélisation

La même expérience a été menée en augmentant le nombre d'attributs de texture sélectionnés. Le tableau 5.3 présente les scores obtenus avec 15 attributs de texture pour chacune des méthodes de sélection de variables testées. Pour la modélisation statistique, 5 attributs sont sélectionnés pour chaque variable forestières et sont ensuite combinés. Ce résultat montre que si les scores sont toujours comparables entre les méthodes pour la ROI 1, pour la ROI 2, la méthode de sélection d'attributs basée sur l'importance de variables des forêts aléatoires est significativement plus performante (amélioration de l'ordre de 5%). Ceci indique que dans un contexte de plus grande dimension, en utilisant un plus grand nombre d'attributs, la méthode de sélection d'attributs basée sur l'importance de variable des forêts aléatoires est à privilégier.

Enfin, si ces premiers résultats nous ont permis de mettre au point notre méthodologie, il faut rappeler que le site de Nezer est un cas simple, avec des peuplements très homogènes. Il est donc apparu essentiel d'enrichir cette comparaison sur le site du Tagon, avec un plus grand nombre d'échantillons et un plus grand nombre d'attributs de texture. Aussi, il est essentiel de prendre en compte toute la complexité de notre problématique, c'est à dire sélectionner les variables à partir d'un ensemble d'attributs image plus grand, issu des 5 bandes

Tableau 5.3 : Performances de classification globale (%) par différentes stratégies de sélection de variables sur deux échantillons de l'image multispectrale Quickbird, avec 15 attributs de texture

	Roi1		Roi2	
	BT	SVM	BT	SVM
<b>RF importance</b>	<b>81.67</b>	<b>81.03</b>	<b>91.29</b>	<b>91.69</b>
LARS stepwise CDHTSP	79.92	81.55	85.83	86.36
Random PRiF CDHTSP	79.2	81.64	85.2	86.95

spectrales disponibles (4 MS et 1 Pan).

### 5.5.2 Expérience complémentaire sur le site du Tagon

Cette expérience a été menée dans le but de déterminer la méthode la plus appropriée pour sélectionner les attributs de texture en amont du classifieur dans un contexte complexe, aussi bien d'un point de vue forestier que d'un point de vue de grande dimensionalité de l'ensemble des attributs images. Pour cela, les 111 placettes correspondant aux mesures in-situ ont été utilisées. L'évaluation des performances a été faite en utilisant les forêts aléatoires avec 100 arbres, 100 placettes pour l'apprentissage et 100 placettes pour le test. Seules les classes de fort (C2 à C5) sont considérées.

Pour la sélection d'attributs basée sur la modélisation statistique, la méthode Random PRiF a été choisie (étant donné qu'elle s'est montrée la plus efficace jusqu'alors). 5 attributs de texture ont été sélectionnés pour chaque variable de la structure forestière et pour chaque résolution. Un ensemble de 50 attributs de texture est donc considéré (25 issus de la résolution multispectrale, 25 issus de la résolution panchromatique). Le résultat est donné sous la forme de matrice de confusion, présentée tableau 5.4.

Pour la sélection d'attributs de texture basée sur l'importance de variables, les forêts aléatoires étant gourmands en espace mémoire, il n'est pas possible de considérer l'ensemble des attributs de texture sur les 111 placettes pour une sélection globale. Nous avons donc testé deux stratégies :

1. Utiliser les 111 placettes pour la sélection d'attributs, en sélectionnant le meilleur paramétrage par attribut et par bande spectrale. Nous avons donc 50 attributs sélectionnés (10 attributs de texture \* 5 bandes spectrales). Le résultat est présenté tableau 5.5.

2. Utiliser un nombre réduit de placettes (40 sur les 111) pour la sélection d'attribut, en sélectionnant de manière globale les 50 premiers attributs de texture issus de l'ensemble complet (tous les paramétrages et toutes les résolutions. Le résultat est présenté tableau 5.6.

### 5.5.2.1 Résultats

Les résultats des deux stratégies utilisant les forêts aléatoires et celle basée sur la modélisation sont présentés dans ce qui suit.

Tableau 5.4 : Matrice de confusion obtenue avec les meilleurs attributs de texture ( $n=50$ ) issus de la sélection de variables Random PRiF pour le site du Tagon, image de Juin.

	C2	C3	C4	C5	Total	Omission%
C2	<b>13899</b>	3587	379	135	18000	22.78
C3	2132	<b>12495</b>	2716	657	18000	30.58
C4	900	1200	<b>12201</b>	3699	18000	32.21
C5	0	444	2191	<b>15365</b>	18000	14.63
Total	16931	17726	17487	19856		
Commission%	17.9	29.51	30.22	22.61		

**Précision globale = 74.94 %**

**Kappa = 0.67**

Tableau 5.5 : Matrice de confusion obtenue avec les meilleurs attributs de texture ( $n=50$ ) obtenus par importance de variables des forêts aléatoires (attribut par attribut et bande spectrale par bande spectrale), image de Juin.

	C2	C3	C4	C5	Total	Omission%
C2	<b>13465</b>	3884	152	499	18000	25.19
C3	2134	<b>12917</b>	1758	1191	18000	28.24
C4	900	2392	<b>9809</b>	4899	18000	45.51
C5	0	592	2311	<b>15097</b>	18000	16.13
Total	16499	19785	14030	21686		
Commission%	18.39	34.71	30.09	30.38		

**Précision globale = 71.23 %**

**Kappa = 0.62**

Ces résultats nous permettent de confirmer ce qui était pressenti avec l'expérience sur le site de Nezer. Si la méthode de sélection de variables basée sur le mod-

Tableau 5.6 : Matrice de confusion obtenue avec les meilleurs attributs de texture ( $n=50$ ) obtenus par importance de variables des forêts aléatoires (de façon globale) pour le site du Tagon, image de Juin.

	C2	C3	C4	C5	Total	Omission%
C2	<b>14416</b>	3174	233	177	18000	24.86
C3	1392	<b>13115</b>	3396	97	18000	27.14
C4	791	864	<b>14821</b>	1524	18000	17.66
C5	1014	2538	2311	<b>14443</b>	18000	19.76
Total	16604	18167	20988	16241		
Commission%	13.18	27.81	28.38	11.07		

**Précision globale = 78.88 %**

**Kappa = 0.72**

élisation fournit des résultats satisfaisants, elle est dépassée (de l'ordre de 4%) par la méthode de sélection de variables basée sur l'importance de variable des forêts aléatoires utilisée de manière globale (et qui utilise donc plus de 50% de données en moins). La stratégie consistant à sélectionner le meilleur paramétrage attribut par attribut n'est pas satisfaisante, elle est dépassée de près de 7.65 % par l'approche globale alors qu'elle est utilisée avec plus de 50% de données supplémentaires.

**5.5.2.1.1 Influence du nombre d'attributs de texture** La figure 5.4 nous permet d'observer l'évolution de la précision globale de classification en fonction du nombre d'attributs de texture. Pour un nombre d'attribut de texture donné, 10 essais de classifications ont été réalisés. Il est à noter que, pour tous les nombres d'attributs de texture testés, l'écartype ne dépasse pas 0,5% de précision globale. Cette courbe montre l'évolution de la précision de classification en fonction du nombre d'attributs de texture utilisés. Un palier est atteint à partir de 50 attributs de texture. Cependant, en pratique, un nombre de 40 attributs de texture semble être un bon compromis entre parcimonie, temps de calcul et précision. Les résultats finaux seront donc réalisés avec 40 attributs de texture.

Cette étude nous a permis de dégager une méthode de sélection d'attributs de texture performante pour la construction de classifieurs, la méthode basée sur l'importance de variable des forêts aléatoires utilisée de façon globale. Cette approche permet d'optimiser la complémentarité entre les attributs de texture

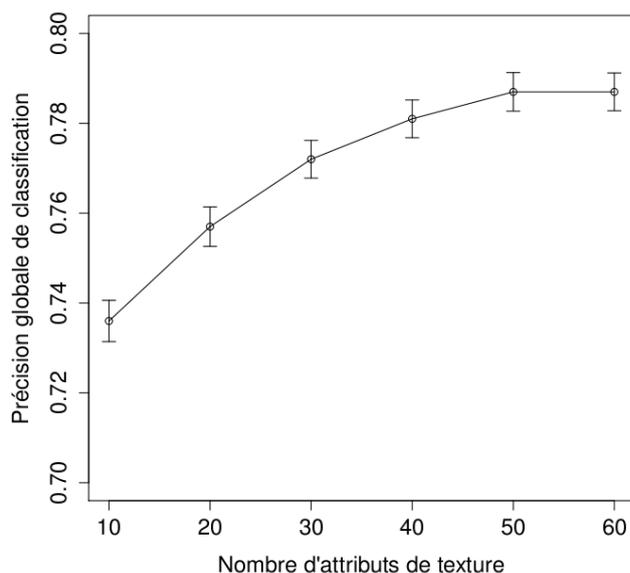


Figure 5.4 : Observation de l'évolution de la précision globale de classification en fonction du nombre d'attributs de texture, image de juin.

issus des deux résolutions, tout en optimisant le paramétrage des attributs de texture de façon automatique. Cette stratégie optimale rejoint la philosophie du Random PRiF utilisé pour l'estimation des variables de la structure forestière, en utilisant la robustesse d'une recherche globale aléatoire. D'un point de vue opérationnel, cette méthode a un avantage supplémentaire, elle est performante en utilisant un nombre réduit d'informations de référence.

## 5.6 Cartographie des structures forestières : vers une solution opérationnelle

### 5.6.1 Une approche hiérarchique

Un premier test sur le site du Tagon (image de juin) nous a amené à considérer une approche hiérarchique, consistant à discriminer dans un premier temps la classe C1 (pas d'arbres visibles), qui correspond aux stades de mutation de la forêt avec un taux de couvert forestier quasi nul, des autres classes C2 à C5 qui correspondent aux stades ultérieurs de développement des peuplements (âge > 3 ans). Cette première classe C1 est donc très différente des autres classes. Cependant elle présente une très forte variabilité dans la composition et le développement de

la végétation (et donc en radiométrie et en texture), ce qui rend sa reconnaissance difficile lorsque l'on cherche à différencier les cinq classes de manière directe. La texture est donc recalculée après avoir masqué la classe 1. Le fait d'éliminer les pixels "non forêt" permet de réduire la variabilité radiométrique des classes et donc d'optimiser la dynamique radiométrique pour les pixels de forêt améliorant ainsi la précision des descripteurs pour la discrimination des classes forestières.

L'expérience suivante illustre cette idée : dans un premier temps, la sélection d'attributs de texture (faite par importance de variables des forêts aléatoires) est réalisée à partir d'attributs de texture calculés sur les pixels de toutes les classes (C1 à C5). Un nombre limité d'attributs (les 15 meilleurs) sont sélectionnés et calculés sur l'image entière avant d'être classifiés par les forêts aléatoires. Cette expérience correspond donc à une stratégie de "classification directe", la matrice de confusion obtenue est présentée tableau 5.7.

Ce résultat nous permet d'observer la difficulté rencontrée par le classifieur lors de cette expérience, la précision globale n'est que de 65%. A partir de la matrice de confusion, on observe un très fort taux d'erreurs, notamment pour les classes C1, C2 et C3, pouvant atteindre 50% d'erreur aussi bien en omission qu'en commission. Ce résultat n'étant pas satisfaisant, comparativement aux scores obtenus pour les classes C2 à C5 précédemment, il a conduit notre réflexion vers une stratégie hiérarchique.

Tableau 5.7 : Matrice de confusion obtenue par classification directe de C1 à C5, image de Juin, 15 attributs de texture.

	C1	C2	C3	C4	C5	Total	Omission%
C1	<b>11038</b>	6789	173	0	0	18000	38.67
C2	5677	<b>9098</b>	2818	225	182	18000	49.45
C3	20	1932	<b>10632</b>	4997	419	18000	40.93
C4	7	892	1390	<b>13503</b>	2208	18000	24.98
C5	54	40	1057	2687	<b>14162</b>	18000	21.32
Total	16796	18751	16070	21412	16971		
Commission%	34.28	51.48	33.83	36.94	16.51		

**Précision globale = 64.95 %**

**Kappa = 0.56**

Dans un second temps, toujours en ne conservant que 15 attributs de texture,

nous avons calculés et sélectionnés les attributs de texture sans prendre en compte la classe C1 d'une part, et d'autre part nous avons regroupé les classes C2 à C5 afin de réaliser une classification C1 contre toutes les autres.

Afin d'obtenir les cartes fusionnées à 5 classes, les pixels classés C1 durant la première étape sont ajoutés à la classification C2 à C5 pour obtenir la carte finale. Cette étape n'est donc pas rigoureusement une classification à 5 classes et une façon appropriée de fusionner ces deux résultats devra être mise au point dans le futur.

Les résultats obtenus sont présentés dans les tableaux 5.8 et 5.9. Ces résultats nous montrent une amélioration très significative de la reconnaissance de la classe C1 (29% d'omission contre 39% auparavant et 16% en commission contre 34% auparavant. De même, la reconnaissance de la classe C2 est améliorée de l'ordre de 30% en omission et 35% en commission. Les précisions de classification sont satisfaisantes avec un indice kappa de 0,68 et une précision globale de 75%. Les taux d'erreurs les plus importants concernent les classes intermédiaires C3 et C4, ce qui était attendu compte tenu de la nature floue de la frontière entre ces classes puisque la structure forestière présente un continuum. Ceci met en évidence la complexité ce type de cartographie.

Ces résultats soulignent l'importance de séparer et de simplifier le problème de classification, en adoptant une approche hiérarchique en deux passes de sélection d'attributs et de classification pour répondre à deux problèmes : la discrimination de la classe C1 (forêt en mutation) dans un premier temps et l'affinage de la discrimination des classes forestières C2 à C5 dans un second temps. Si ce processus augmente fortement la complexité des calculs, il est justifié par une amélioration globale de la précision de l'ordre de 10% ce qui n'est pas négligeable. De plus, d'un point de vue thématique, la discrimination C1 contre les autres classes a un fort intérêt puisqu'elle correspond à la cartographie des surfaces non boisées et boisées.

Les cartes de classification associées sont présentées dans la section suivante.

### **5.6.2 Utilisation de la marge d'ensemble :**

Les sections précédentes nous ont permis de justifier notre approche hiérarchique et de mettre au point notre stratégie de sélection d'attributs de texture pour la cartographie des types de structure forestière. Nous sommes donc en mesure

Tableau 5.8 : Matrice de confusion obtenue C2-C5, image de Juin, 15 attributs de texture.

	C2	C3	C4	C5	Total	Omission%
C2	<b>14595</b>	2514	804	87	18000	18.92
C3	1488	<b>12224</b>	4198	90	18000	32.08
C4	828	2054	<b>13508</b>	1610	18000	24.95
C5	1	1682	2115	<b>14202</b>	18000	21.10
Total	16912	18474	20625	15989		
Commission%	13.70	33.83	34.50	11.18		

**Précision globale = 75.73 %**

**Kappa = 0.68**

Tableau 5.9 : Matrice de confusion obtenue C1 VS autres image de Juin, 15 attributs de texture.

	C1	Autres	Total	Omission%
C1	<b>12757</b>	5243	18000	29.13
Autres	2460	<b>69540</b>	72000	3.42
Total	15217	74783		
Commission%	16.17	7.01		

**Précision globale = 91.44 %**

**Kappa = 0.72**

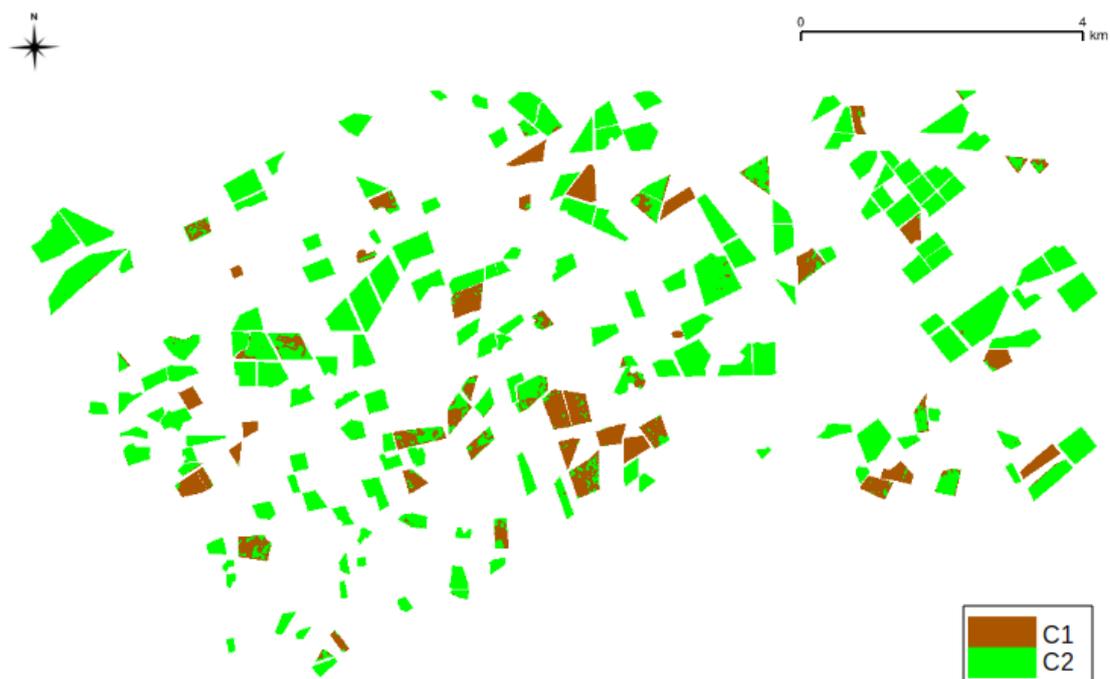
de produire des cartographies des types de structure forestières. Cette section présente un intérêt thématique et propose une utilisation pratique de l'information de marge d'ensemble fournie par les forêts aléatoires.

### 5.6.2.1 Les cartes de marge.

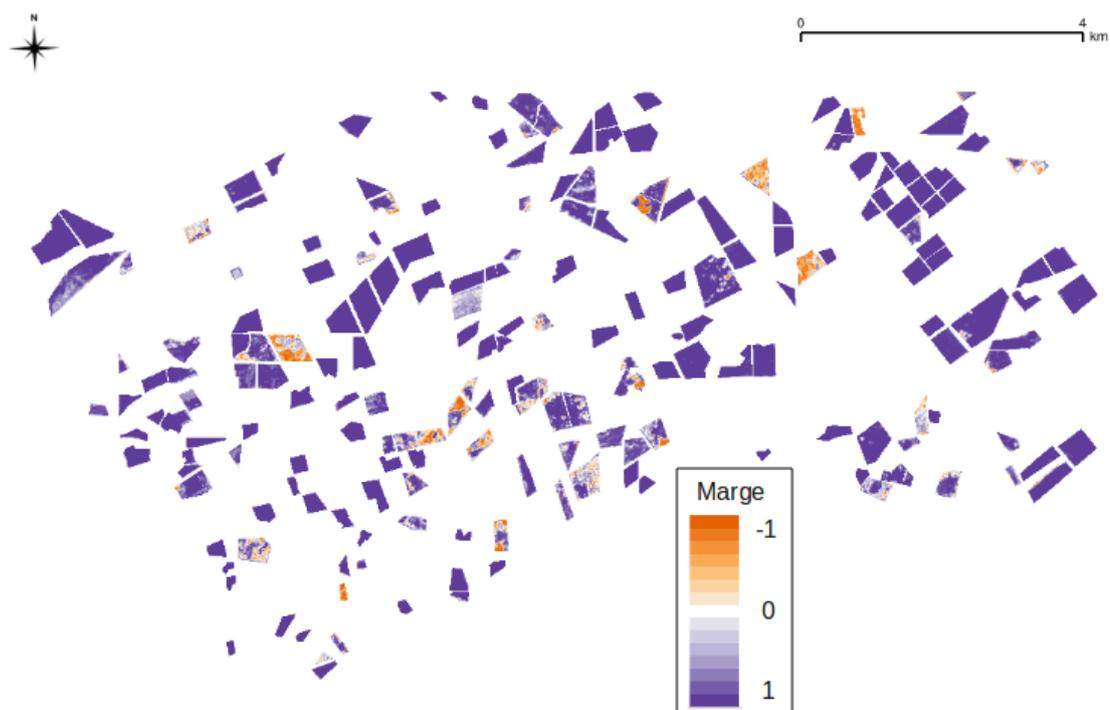
Ces cartes et les résultats de classification associées sont présentées figure 5.5 pour la classification C1 versus Autres et figure 5.6 pour la classification des classes C2 à C5.

En effet, en observant les cartes de marges, certaines régions, ou parcelles sont apparues avec des marges très significativement négatives (rouges sur les images). Ceci indique d'une part que le classifieur est confiant sur la classe prédite (marge proche de 1 en valeur absolue) mais aussi que la classe prédite ne correspond pas à la classe de référence (marge négative). Pour la classification C1 versus Autres, la figure 5.5 montre que la classification réalisée est très fiable (marges

majoritairement proches de 1). Seules quelques zones sont orangées, elles correspondent à des parcelles non boisées mais fortement végétalisées, ce qui explique la confusion du classifieur. Pour la classification C2 à C5, la classification est moins fiable, notamment sur les bords de certaines parcelles. Aussi un nombre relativement important de parcelles sont "mouchetées" ou "pâles" ce qui indique que ces peuplements sont à la frontière entre deux classes, présentant une forte variabilité intra-parcellaire.

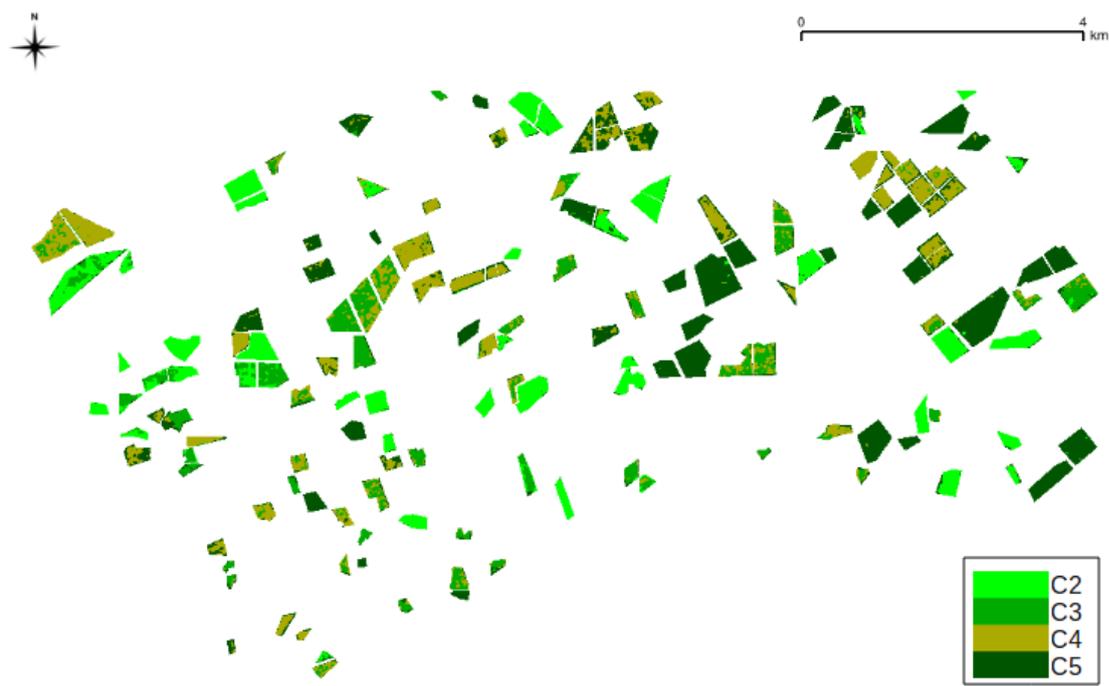


(a) Classification C1 versus C2 (Autres), image de juin.

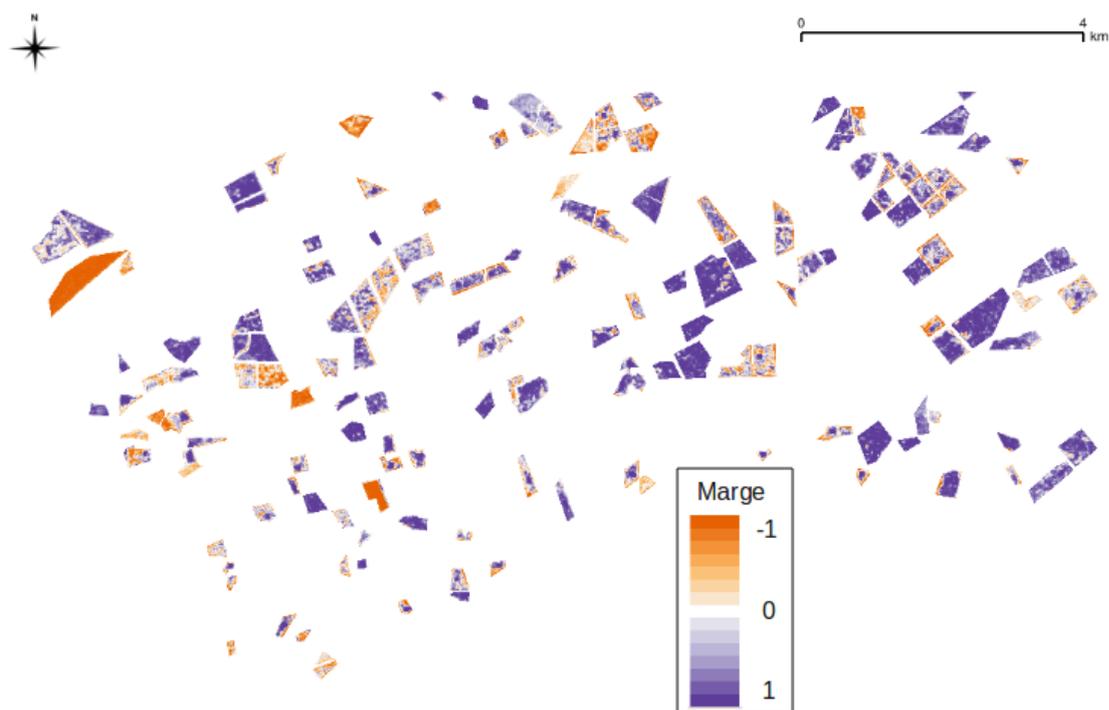


(b) Carte de marge correspondante.

Figure 5.5 : Classification C1 versus C2 (Autres) et carte de marge associée, image de juin.



(a) Classification C2 à C5.



(b) Carte de marge correspondante.

Figure 5.6 : Classification C2 à C5 et carte de marge associée, image de juin.

### 5.6.2.2 Détection des erreurs d'étiquetage

Les cartes de marge, permettent ainsi d'identifier les régions qui sont complexes à cartographier par le classifieur. Dans ce travail, l'étiquetage des données de référence étant fait par photo-interprétation, une certaine incertitude sur la base de données de référence peut exister. D'autre part, l'étiquetage au niveau de la parcelle peut être complexe en cas de forte variabilité intra-parcellaire.

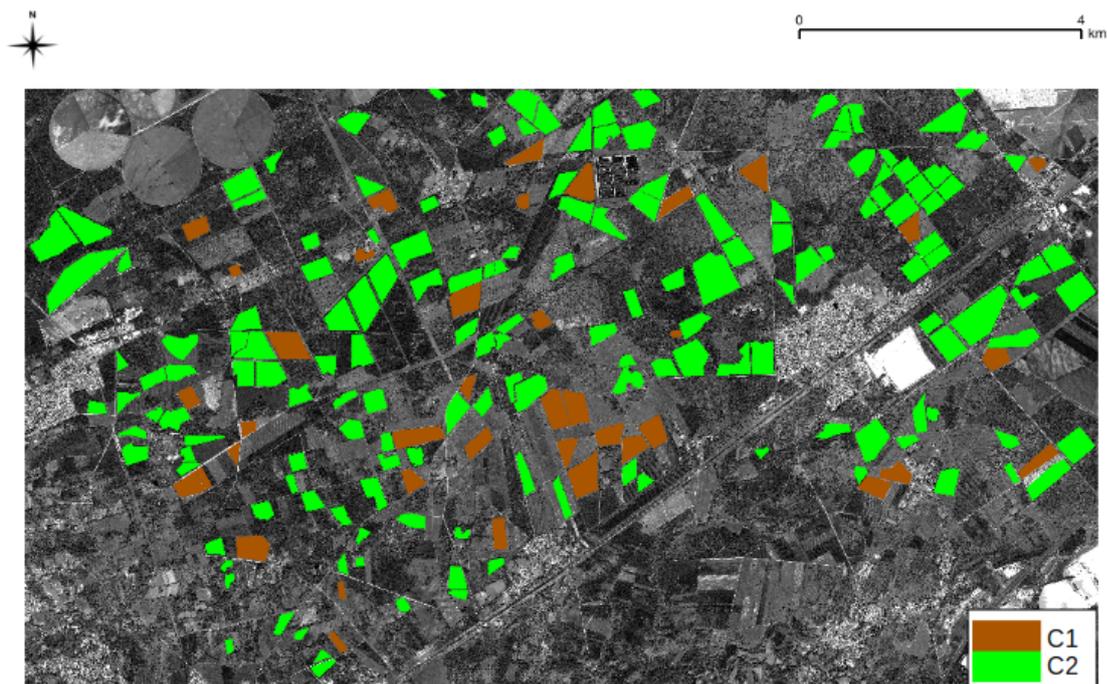
L'idée est alors apparue d'utiliser les cartes de marge, non seulement pour localiser les difficultés du classifieur, mais aussi pour détecter d'éventuelles erreurs d'étiquetage commises par l'opérateur humain. En effet, les faibles marges correspondent à des pixels mal classés avec une forte confiance du classifieur et donc à des probables erreurs d'étiquetage.

Pour la détection et la correction des erreurs d'étiquetage, nous avons choisi de les réaliser à l'échelle de la parcelle pour minimiser le coût d'intervention de l'opérateur humain. Pour chaque placette de la base de test, la marge moyenne est calculée. Un seuil d'alarme pour une marge moyenne de -0,3 a été utilisé pour identifier automatiquement les parcelles problématiques et donc les éventuelles erreurs d'étiquetage.

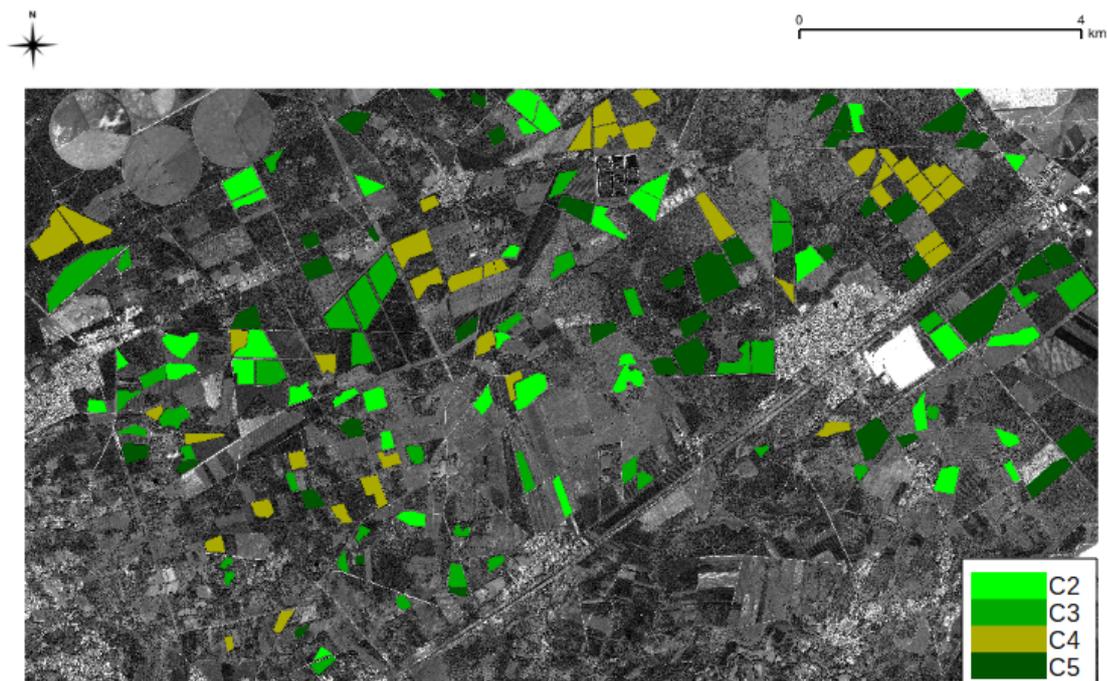
Un contrôle par un opérateur humain est alors effectué et une correction manuelle de l'étiquetage est réalisée, si besoin, par photo-interprétation.

*Tableau 5.10 : Résumé des alarmes et corrections d'étiquetage détectés avec un seuil de marge moyenne à -0.3.*

Numéro	Avant	Après	Commentaires
<b>174</b>	<b>C4</b>	<b>C3</b>	<b>Erreur évidente</b>
172	C3	C3	Classe d'origine validée
164	C5	C5	Placette dans une zone d'erreurs, classe d'origine validée
49	C5	C5	Classe d'origine validée, peuplement très dense, faible espacement
<b>145</b>	<b>C5</b>	<b>C4</b>	<b>Classe 4 avérée, frontière entre C4 et C5</b>
<b>83</b>	<b>C5</b>	<b>C4</b>	<b>Classe 4 avérée, frontière entre C4 et C5</b>
<b>165</b>	<b>C4</b>	<b>C5</b>	<b>Parcelle très complexe, peuplement très mature</b>
8	C2	C2	Classifieur perturbé mais classe d'origine validée
<b>220</b>	<b>C2</b>	<b>C3</b>	<b>Classe 3 avérée, frontière entre C2 et C3</b>
<b>222</b>	<b>C2</b>	<b>C3</b>	<b>Classe 3 avérée, frontière entre C2 et C3, nombreuses trouées</b>
<b>42</b>	<b>C3</b>	<b>4</b>	<b>Classe 4 avérée, frontière entre C3 et C4</b>



(a)Echantillons de référence (C1 versus Autres).



(b)Echantillons de référence, C2 à C5.

Figure 5.7 : Echantillons de référence après corrections d'étiquetage, fond : image panchromatique du mois de juin.

Pour la classification C1 versus Autres, seules deux placettes étaient alarmantes et une seule d'entre elle était mal étiquetée. Pour la classification C2 à

C5, le tableau 5.10 reporte les corrections d'étiquetage ainsi que les observations correspondantes. 11 placettes étaient alarmantes et 7 d'entre elles ont été corrigées, Le total des erreurs d'étiquetage s'élevait à 6% de l'ensemble de la base de référence. Ceci confirme l'incertitude présente dans la base de référence, constituée à partir d'un milieu forestier assez complexe.

Les images de références ainsi corrigées ont ensuite été utilisées pour évaluer les cartographies finales, elles sont présentées figure 5.7. Les précisions de classifications finales sont présentées dans la section suivante. On peut alors envisager une utilisation opérationnelle de ce système d'alarme par la marge d'ensemble pour détecter les erreurs d'étiquetage dans une base de données de référence et ainsi améliorer à la fois le travail de photo-interprétation mais aussi les résultats de cartographie pour des milieux complexes tels que la forêt.

### **5.6.2.3 La marge : support d'interprétation.**

Les figures 5.9 et 5.8 permettent de visualiser de façon plus précise les cartes obtenues.

La figure 5.8 permet de visualiser les difficultés de la tâche de classification. Pour la parcelle centrale de l'image, étiquetée comme classe 4, on peut observer sur l'image panchromatique associée une forte variabilité intra parcellaire. Certaines zones présentent une forte densité d'arbres tandis que d'autres, probablement suite à des éclaircies, présentent une densité plus faible. On peut alors observer sur la carte de marge trois types de zones pour cette parcelle. Les zones bien "bleues", c'est à dire des zones pour lesquelles le classifieur est confiant et la classe prédite est la classe de référence. Des zones sont ensuite très pâles, le classifieur indique ici une faible marge et donc une faible confiance, nous sommes alors en présence de zones qui sont en frontière de classes. Enfin, les zones rouges indiquent que le classifieur est confiant et que la classe prédite n'est pas la bonne, ces zones sont majoritairement localisées sur les bords de la parcelle. On peut observer sur la carte de classification que la classe prédite pour ces zones est la classe 5. En observant l'image panchromatique, ces zones correspondent à des densités plus faibles que les densités au centre de la parcelle, ceci indiquant la pertinence de cette prédiction et le potentiel de notre méthodologie pour repérer des variations intra-parcellaires. Enfin, une petite zone de classe 3 a été prédite au centre de cette parcelle, il semble en effet, au regard de l'image panchromatique,

que cette zone corresponde à une densité d'arbres plus importante que pour le reste de la parcelle. La pertinence de cette prédiction est à relativiser au regard de la carte de marge qui indique une faible confiance pour cette zone.

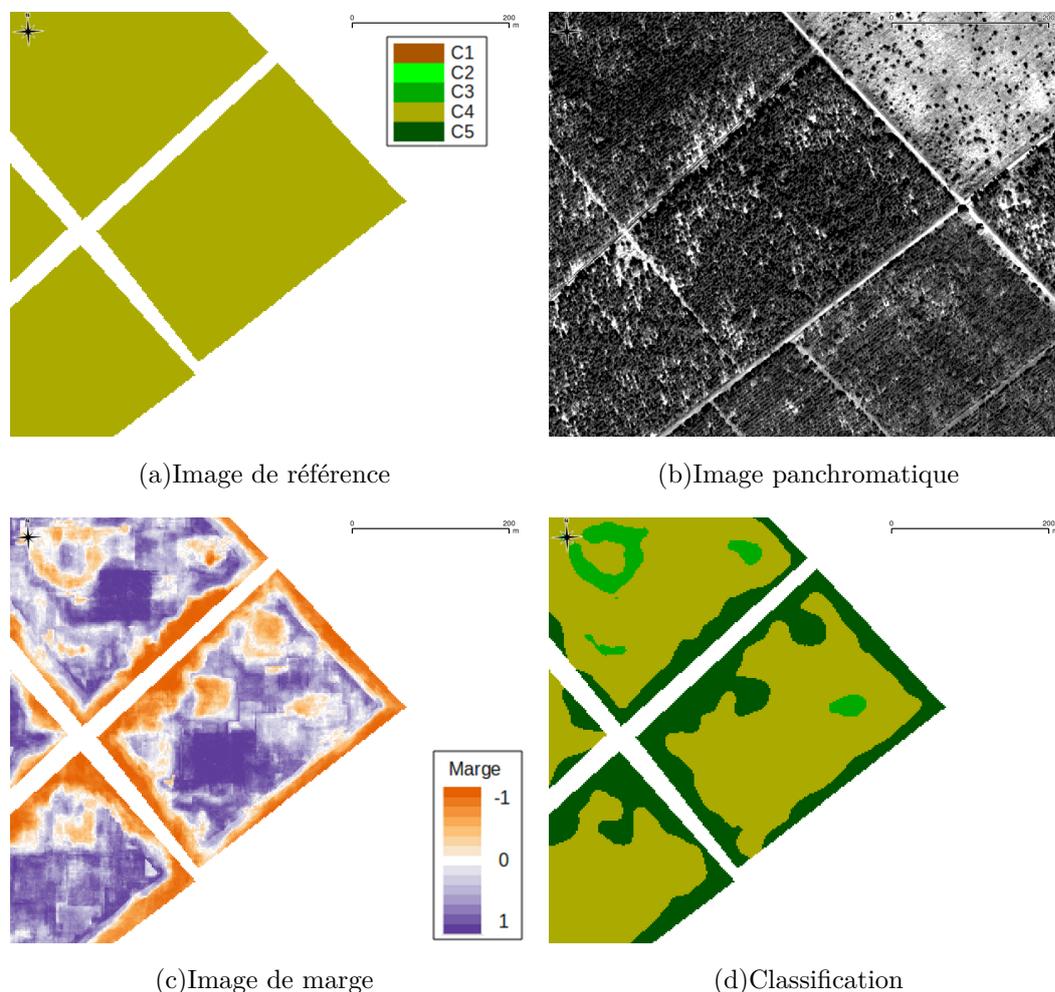


Figure 5.8 : Exemple d'une parcelle complexe et de certaines difficultés rencontrées par le classifieur, image Pléiades du mois de juin.

La figure 5.9 démontre le potentiel de la méthodologie proposée pour détecter des trouées au sein d'un peuplement. Si l'image de référence indique une classe 5 pour l'ensemble de la parcelle, l'image panchromatique montre une zone de trouée importante. Cette zone est bien localisée dans la carte de marge avec une tache rouge, cette tache indique qu'une zone a été prédite dans une autre classe que la classe de référence et cela avec une forte confiance. La carte de classification obtenue met bien en évidence cette trouée dans ce peuplement mature. Il est intéressant d'observer que dans cette parcelle à forte densité, les espacements nor-

maux entre les arbres peuvent être importants mais ils ne sont pas détectés comme étant des trouées. Un cas limite est observable en bas à droite de cette parcelle, une petite tache différente de la carte de référence est prédite, la carte de marge indique une faible confiance pour cette zone de la classification. L'observation de l'image panchromatique à cet endroit montre bien que nous sommes à la limite entre une trouée à détecter et un espacement normal entre les arbres pour une parcelle à ce stade de maturité.

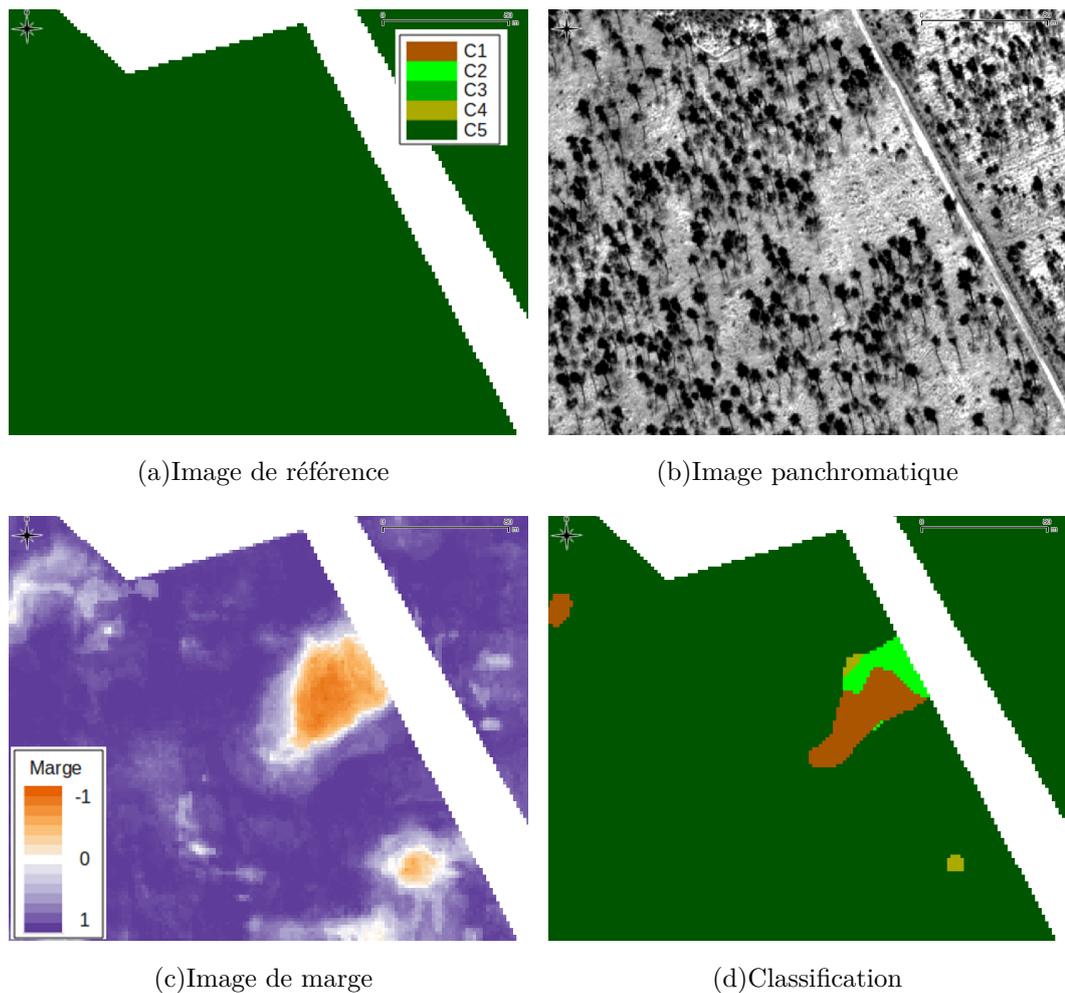


Figure 5.9 : Exemple de localisation d'une trouée dans un peuplement mature (classe 5), image Pléiades du mois de juin.

L'observation de ces résultats de classification montre le fort potentiel des images THR et de la méthodologie développée pour décrire finement la structure forestière en détectant des variations fines au sein des peuplements.

### 5.6.3 Evaluation multi-dates

En utilisant la base de données de référence corrigée, cette section présente les cartographies finales des types de structure forestière pour les trois images Pléiades (Juin, Août et février). Pour chacune des dates sont présentées les matrices de confusion pour la première étape de notre stratégie hiérarchique (C1 Vs. Autres), pour la seconde étape (C2 à C5) ainsi qu'une estimation de la précision globale obtenue sur les 5 classes.

Les précisions globales obtenues sont très satisfaisantes compte tenu de la complexité du problème et de la grande hétérogénéité présente sur le site du Tagon. Pour les trois dates, notre méthodologie a permis de trouver, de manière automatique, les bons attributs de texture et ainsi réaliser des classifications de qualité. La reproductibilité des bons scores, sur trois jeux de données, à trois dates différentes, souligne la robustesse de notre approche.

Grâce à l'automatisation des procédures, l'intérêt est de ne faire aucune hypothèse a priori sur la relation entre structure forestière et texture d'image, qui peut être très dépendante des conditions d'observation satellitaire (saison qui détermine la hauteur du soleil et le stade phénologique, angles de visée, par exemple).

Pour les classifications de C1 Vs. Autres, les précisions globales sont de 94%, 95% et 98,5% pour Juin, Août et Février respectivement. L'amélioration de ce score en Février peut être expliqué par le fait que la végétation du sois bois est sèche, rendant la discrimination surface boisée Vs. surface non boisée plus évidente. On peut alors conseiller, pour ce type d'applications, de travailler avec des images d'hiver.

Pour les cartographies des classes C2 à C5, les précisions globales sont de 86%, 83,4% et 86% pour Juin, Août et Février respectivement. Enfin, pour les cartographies finales (C1 à C5), les précisions globales sont de 82%, 81,6% et 83,4% pour Juin, Août et Février respectivement. Ces scores sont très satisfaisants et encourageants, ils permettent d'envisager une utilisation opérationnelle de la méthodologie proposée pour la gestion de la ressource forestière.

Tableau 5.11 : Matrice de confusion obtenue pour C1 VS. Autres, image de Juin.

	C1	Autres	Total	Omission%
C1	<b>13313</b>	3787	17100	22.14
Autres	1984	<b>70016</b>	72000	2.75
Total	15297	73803		
Commission%	12.96	5.13		

**Précision globale = 94 %**

**Kappa = 0.78**

Tableau 5.12 : Matrice de confusion obtenue pour C2-C5, image de Juin.

	C2	C3	C4	C5	Total	Omission%
C2	<b>14508</b>	1115	474	103	16200	10.44
C3	1474	<b>15487</b>	2742	97	19800	21.78
C4	0	774	<b>16919</b>	1207	18900	10.48
C5	0	969	1026	<b>15105</b>	17100	11.66
Total	15982	18345	21161	16512		
Commission%	9.22	15.58	20.04	8.52		

**Précision globale = 86 %**

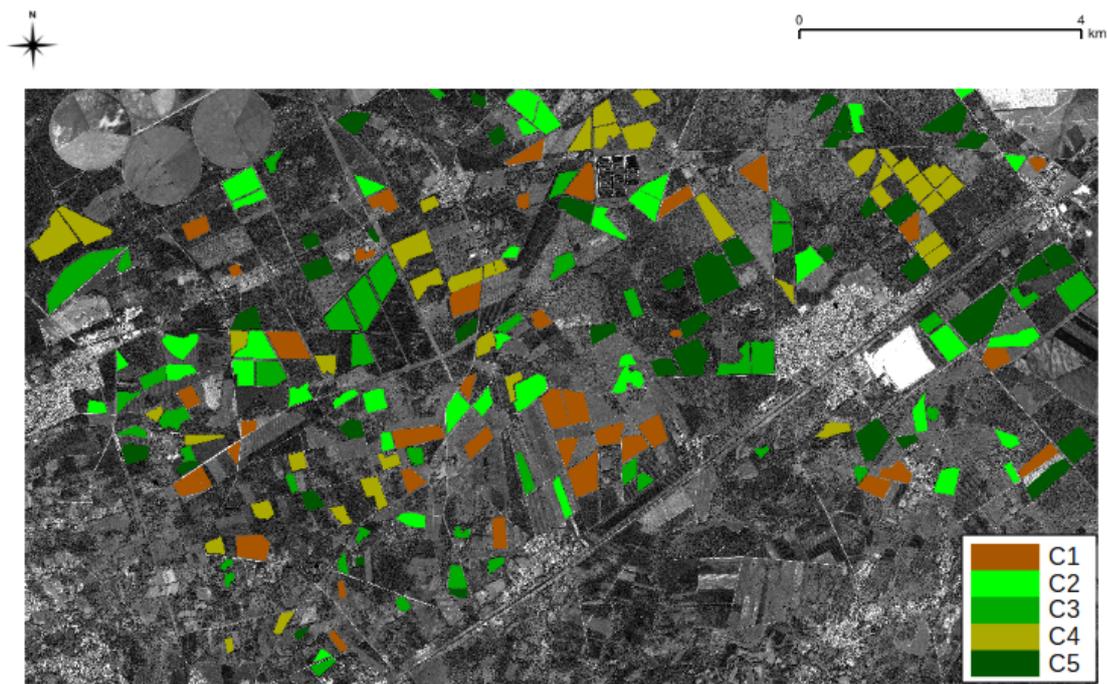
**Kappa = 0.81**

Tableau 5.13 : Matrice de confusion obtenue pour C1-C5, image de Juin.

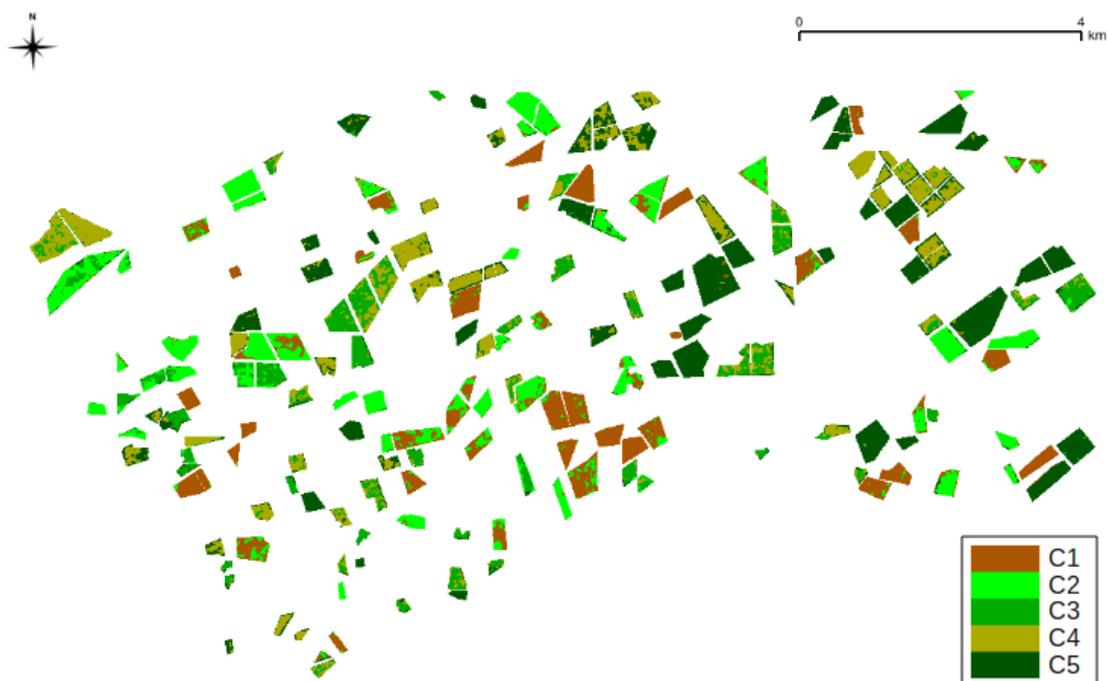
	C1	C2	C3	C4	C5	Total	Omission%
C1	<b>13313</b>	3787	0	0	0	17100	22.14
C2	1983	<b>12657</b>	1020	474	66	16200	21.87
C3	1	1473	<b>15487</b>	2742	97	19800	21.78
C4	0	0	774	<b>16919</b>	1207	18900	10.48
C5	0	0	969	1026	<b>15105</b>	17100	11.67
Total	15297	17917	18250	21161	16475		
Commission%	12.97	29.35	15.13	20.05	8.32		

**Précision globale = 82 %**

**Kappa = 0.78**



(a)Echantillons de référence C1 à C5.



(b)Classification C1 à C5)

Figure 5.10 : Classification finale, mois de juin.

Tableau 5.14 : Matrice de confusion obtenue pour C1 Vs. Autres image d'Août.

	C1	Autres	Total	Omission%
C1	<b>13186</b>	3914	17100	22.89
Autres	508	<b>71492</b>	72000	0.70
Total	13694	75406		
Commission%	3.70	5.19		

**Précision globale = 95 %**

**Kappa = 0.82**

Tableau 5.15 : Matrice de confusion obtenue pour C2-C5, image d'Août.

	C2	C3	C4	C5	Total	Omission%
C2	<b>13634</b>	2353	189	24	16200	15.84
C3	3139	<b>15137</b>	1232	292	19800	23.55
C4	0	2458	<b>16006</b>	436	18900	15.31
C5	0	0	1848	<b>15252</b>	17100	10.80
Total	16773	19948	19275	16004		
Commission%	18.71	24.11	16.95	4.70		

**Précision globale = 83.37 %**

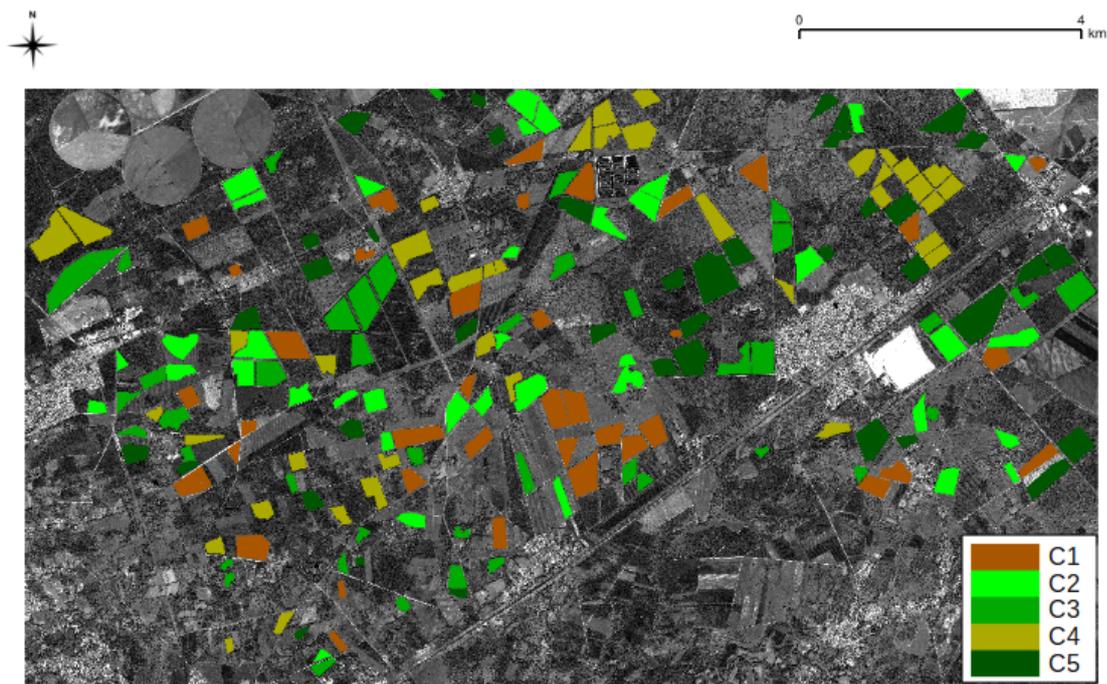
**Kappa = 0.78**

Tableau 5.16 : Matrice de confusion obtenue pour C1-C5, image d'Août.

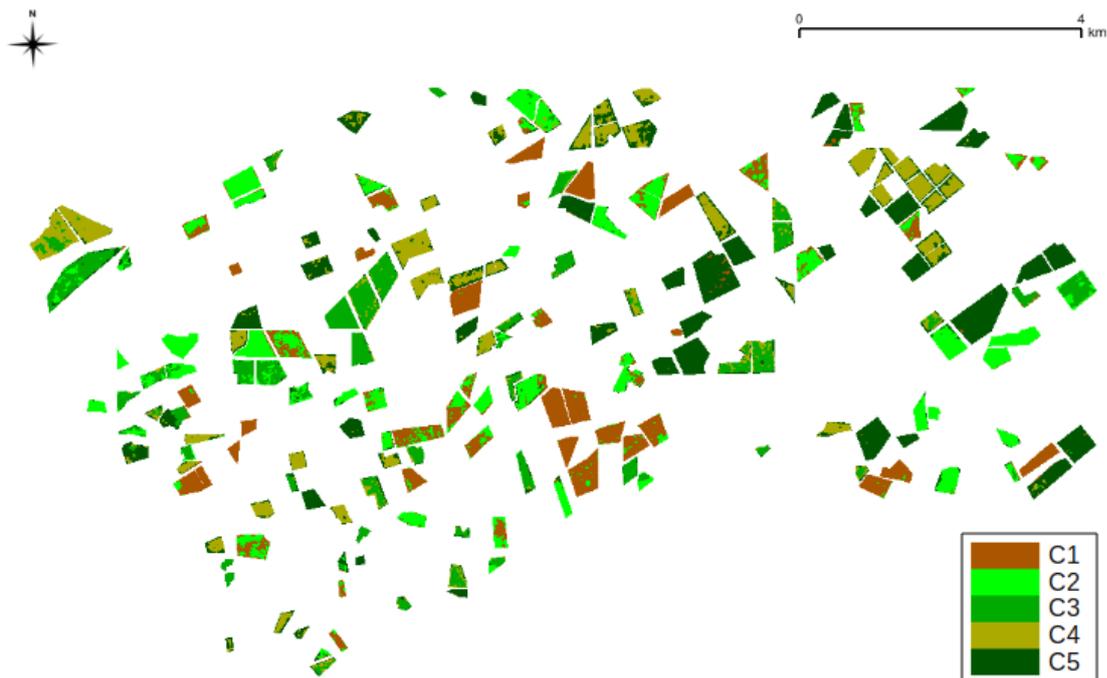
	C1	C2	C3	C4	C5	Total	Omission%
C1	<b>13186</b>	3914	0	0	0	17100	22.89
C2	504	<b>13130</b>	2353	189	24	16200	18.95
C3	2	3139	<b>15137</b>	1232	290	19800	23.55
C4	0	0	2458	<b>16006</b>	436	18900	15.31
C5	2	0	0	1848	<b>15250</b>	17100	10.81
Total	13694	20183	19948	19275	16000		
Commission%	3.71	34.95	24.12	16.96	4.69		

**Précision globale = 81.60 %**

**Kappa = 0.77**



(a)Echantillons de référence C1 à C5.



(b)Classification C1 à C5)

Figure 5.11 : Classification finale, mois d'Août.

Tableau 5.17 : Matrice de confusion obtenue pour C1 Vs. Autres image de février.

	C1	Autres	Total	Omission%
C1	<b>16510</b>	590	17100	3.45
Autres	810	<b>71190</b>	72000	1.13
Total	17320	71780		
Commission%	4.68	0.82		

**Précision globale = 98.43 %**

**Kappa = 0.95**

Tableau 5.18 : Matrice de confusion obtenue pour C2-C5, image de février.

	C2	C3	C4	C5	Total	Omission%
C2	<b>13088</b>	3112	0	0	16200	19.21
C3	2766	<b>17001</b>	10	23	19800	14.14
C4	0	1856	<b>13292</b>	3752	18900	29.67
C5	0	424	1465	<b>15211</b>	17100	11.05
Total	15854	22393	14767	18986		
Commission%	17.44	24.8	10.00	19.88		

**Précision globale = 86 %**

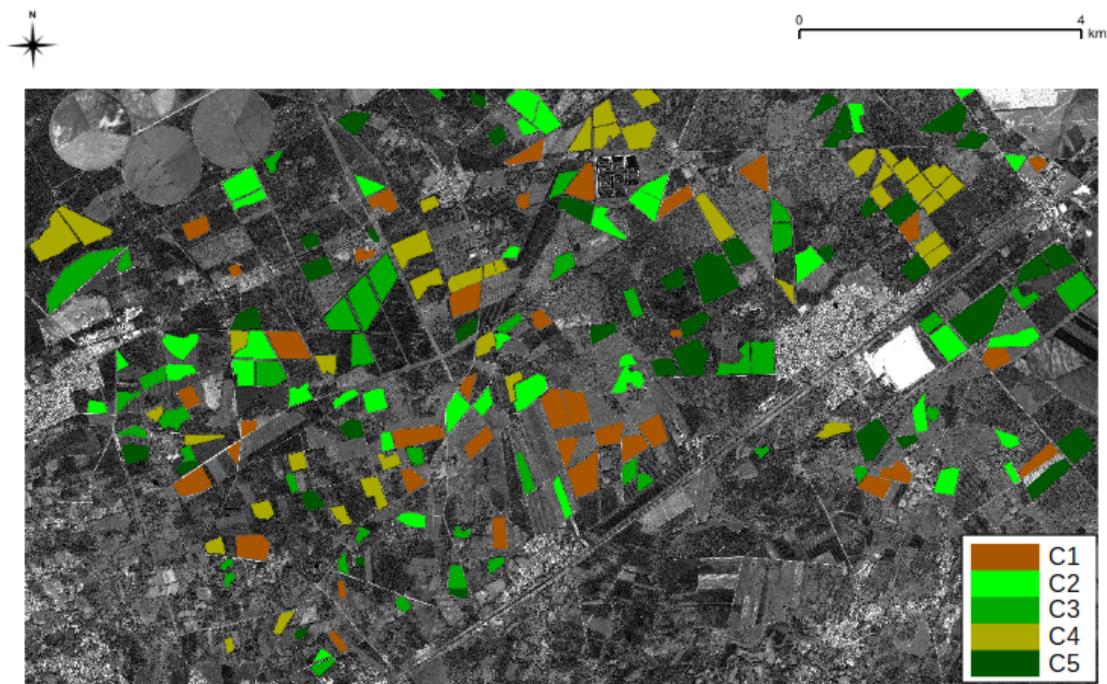
**Kappa = 0.81**

Tableau 5.19 : Matrice de confusion obtenue pour C1-C5, image de février.

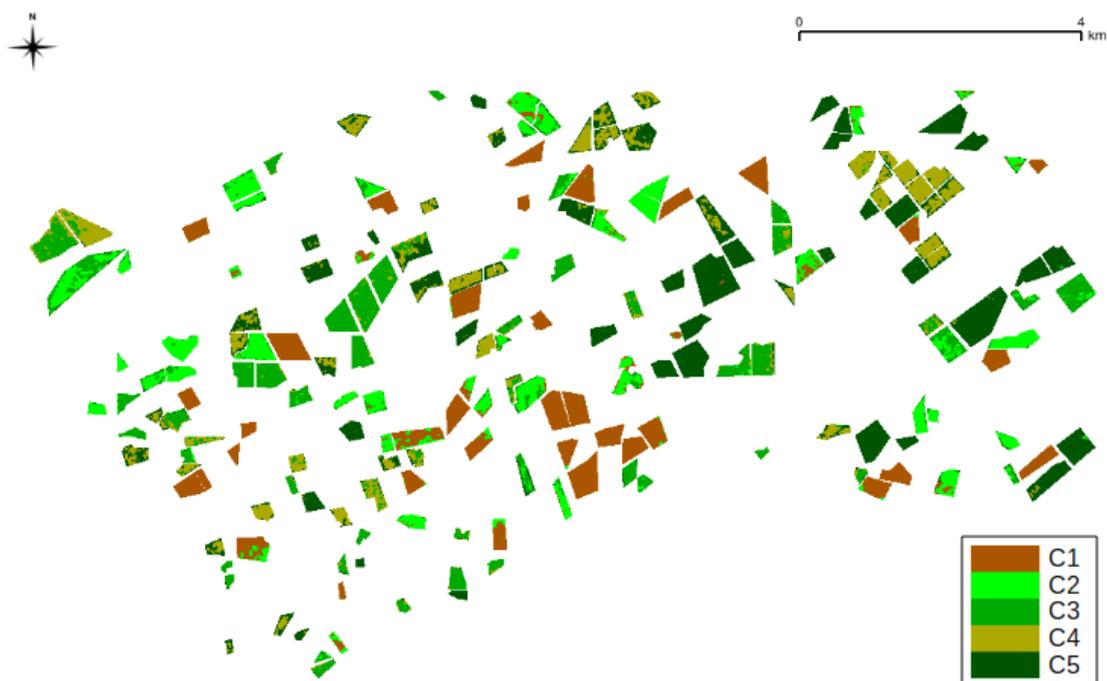
	C1	C2	C3	C4	C5	Total	Omission%
C1	<b>16510</b>	590	0	0	0	17100	3.45
C2	810	<b>12278</b>	3112	0	0	16200	24.21
C3	0	2766	<b>17001</b>	10	23	19800	14.13
C4	0	0	1856	<b>13292</b>	3752	18900	29.67
C5	0	0	424	1465	<b>15211</b>	17100	11.05
Total	17320	15634	22393	14767	18986		
Commission%	4.68	21.47	24.08	10.00	19.88		

**Précision globale = 83.38 %**

**Kappa = 0.79**



(a)Echantillons de référence C1 à C5.



(b)Classification C1 à C5)

Figure 5.12 : Classification finale, mois de Février.

## 5.7 Conclusions

Ce chapitre nous a permis de développer et de tester une stratégie pour la cartographie des types de structure forestières. Une approche hiérarchique est proposée et validée. Une première étape permet de discriminer les peuplements forestiers en mutation (coupes rases, reboisements en cours ou très récents) des autres stades forestiers. Un taux de précision globale de l'ordre de 94% est obtenu. Cette étape est aussi intéressante dans la perspective de la détection de changements (coupe rase, nettoyage ou labour indicateurs d'un reboisement, ou encore dégâts de tempête). La classification en quatre grandes classes allant du stade de jeune peuplement ( $> 3$  ans) à celui de peuplement mature ( $50$  ans) a atteint un score de précision globale de l'ordre de 85%. Ces résultats, établis en supposant que la localisation de la forêt de pin maritime et les limites de peuplement sont connues, sont très encourageants quant à des perspectives opérationnelles. En effet, à titre de comparaison, la carte des classes d'âge produite occasionnellement par l'IFN (service de l'Inventaire Forestier National) sur le département des Landes (BD Forêt Version 1 de l'IGN) par photointerprétation de prises de vue aérienne (résolution  $50$  cm) différencie 3 classes d'âge dans la futaie pure de pin maritime.

La sélection des attributs de texture pour la classification est faite par importance de variables des forêts aléatoires. Cette approche nous permet, en utilisant un nombre réduit de données de référence, de sélectionner et de combiner des attributs de texture issus des deux résolutions (panchromatique et multispectale) permettant ainsi une approche multi-échelle et multi-résolution. Cette approche globale s'est montrée la plus performante parmi les approches testées.

L'utilisation des forêts aléatoires nous a permis de dégager une possible utilisation de l'information de la marge d'ensemble : la correction des erreurs d'étiquetage dans la base de donnée de référence. Cette notion est très intéressante pour les problèmes de classification sur des surfaces complexes, ou de manière générale pour les cas où les données de référence sont construites par un opérateur humain. La marge d'ensemble permet donc à l'utilisateur de vérifier la qualité des données en plus de fournir une information sur la fiabilité de la classification et sur la localisation des difficultés rencontrées par le classifieur.

La robustesse de la méthodologie proposée et la reproductibilité des hauts niveaux de précision atteints sont confirmés avec trois images Pléiades, acquises

dans des conditions très différentes : en hiver (février 2013) avec une hauteur solaire moins élevée et une végétation du sous-bois sèche et en été (juin et août 2012) avec une hauteur solaire plus élevée et une végétation du sous-bois bien verte.

L'application aux images multi-annuelles pourra permettre d'évaluer sur le massif landais la possibilité de cartographier des changements tels que des coupes rases, le mitage urbain ou les dégâts de tempête.

# Chapitre 6

## Conclusions et perspectives

L'étude et la surveillance des écosystèmes forestiers est une priorité scientifique, politique et socio-économique. La télédétection et plus particulièrement l'imagerie satellitaire représente une solution peu coûteuse et prometteuse. Depuis une dizaine d'années, le nombre de données satellitaires optiques à Très Haute Résolution (THR) spatiale ne cesse d'augmenter, fournissant des données de précision à divers utilisateurs. L'objectif principal de cette étude a été donc d'évaluer le potentiel des images satellitaires à Très Haute résolution spatiale (et plus particulièrement des images Pléiades) pour répondre au fort besoin d'inventaires fins et spatialisés de la ressource forestière.

Ce travail de thèse s'est focalisé sur deux objectifs complémentaires. Le premier vise l'estimation de variables de structure forestière (grandeurs dendrométriques telles que diamètre des couronnes, hauteur, diamètre des troncs, densité ou espacement des arbres) via une modélisation de leur relation avec des attributs de texture. Les erreurs d'estimation n'étant pas trop fortes devant l'étendue de ces variables, il est alors possible d'identifier plusieurs classes de structure forestière couvrant tout le cycle de développement des peuplements. Le second objectif vise alors à identifier différents stades de développement des peuplements (du renouvellement à la maturité) à l'aide d'une classification d'image au niveau pixel de façon à produire une information forestière spatialisée la plus fine possible. Le bilan de ce travail et les perspectives sont présentés dans ce qui suit.

## 6.1 Analyse de texture

Pour répondre aux deux objectifs principaux de ce travail de thèse, à savoir l'estimation des variables forestières et la cartographie des types de structure forestière à partir de la texture des images à très haute résolution spatiale, nous avons choisi d'utiliser les attributs de textures issus des matrices de co-occurrence. La principale limite de cet outil d'analyse de texture étant le grand nombre de paramètres spatiaux à déterminer, la contribution majeure de ce travail de thèse est de proposer une méthode automatique qui, balayant une vaste gamme de paramétrages spatiaux, permet d'optimiser ces paramétrages. L'idée principale est donc de calculer un grand nombre d'attributs de texture sur un ensemble réduit d'échantillons d'images avant d'en sélectionner un certain nombre par une méthode appropriée à chacun de nos objectifs. De plus, nous avons montré que calculer les attributs de texture sur l'ensemble des données disponibles (aux deux résolutions et sur toutes les bandes spectrales) avant de sélectionner un sous ensemble permet, sans a priori, d'optimiser l'utilisation de la donnée en fournissant des descripteurs multi-résolution et multi-échelle.

## 6.2 Estimation des variables forestières

Afin d'explorer le potentiel des images THR pour la caractérisation de la structure forestière, des régressions linéaires multiples entre chaque variable de structure forestière et un ensemble d'attributs de texture optimal sont appliquées. Nous avons montré les fortes relations entre la texture des images et les variables de la structure forestière. Nous avons pu établir des modèles de prédiction performants pour chacune de ces variables forestières. Différentes méthodes de sélection de variables pour la régression linéaire multiple ont été testées, une attention particulière a été portée au problème de la colinéarité entre les variables (i.e. les attributs image). Une contribution importante a été apportée par la méthode de sélection de variables Random PRiF développée durant cette thèse. Cette nouvelle méthode a permis d'obtenir les modèles les plus performants en terme de prédiction en s'assurant d'éviter le problème de la colinéarité entre les variables. Cette méthode a été appliquée avec succès sur les deux sites d'études, avec différentes configurations (angles de visée, angles solaires, saison). Les évaluations des performances, menées notamment sur les 3 images Pléiades, confirment la robustesse de la méthodologie. Les performances de prédiction des différentes variables forestières étudiées sont très bonnes.

## 6.3 Cartographie des structures forestières

Pour la cartographie des types de structure forestière fondée sur une classification d'image, une stratégie à visée opérationnelle a été développée. Nous proposons une approche hiérarchique qui consiste dans un premier temps à discriminer les différents stades de mutation de la forêt (coupes rases, reboisements en cours ou récents) des stades forestiers à proprement parler (qui correspondent aux peuplements dont l'âge est supérieur à environ trois ans). Une fois cette première stratification réalisée, la seconde étape de cette approche hiérarchique se concentre sur la discrimination des stades de développement des peuplements forestiers. Cette approche est basée sur une sélection en amont d'un ensemble d'attributs de texture issus des deux résolutions. Les forêts aléatoires (Random Forest (RF)), classifieur d'ensemble, sont utilisés pour la classification et la sélection de variables. La sélection des attributs de texture est ici réalisée grâce à l'importance de variables obtenue par les forêts aléatoires. Un des intérêts principaux de ce classifieur est le concept de marge d'ensemble qui mesure la difficulté du classifieur à classer une instance. Ces cartes de marge nous ont conduits à un apport thématique important : la détection des erreurs d'étiquetage dans la base de données de référence. Elles sont aussi très utiles pour l'interprétation des cartes de classification obtenues et constituent en pratique des cartes d'alerte pour un opérateur humain permettant de réduire le coût de son intervention pour le post-traitement des cartographies. Les résultats, établis en supposant que la localisation de la forêt de pin maritime et les limites de peuplement sont connues, sont très encourageants quant à des perspectives opérationnelles.

## 6.4 Perspectives

Pour conclure le travail sur les attributs issus des matrices de cooccurrence, il serait intéressant d'inclure tous les attributs d'Haralick (14) et non seulement les huit attributs les plus utilisés en littérature afin d'en observer les possibles performances dans le cadre de combinaison d'attributs multirésolution et multiéchelle.

Enfin, cette étude s'est concentrée sur l'utilisation des attributs de texture issus des matrices de co-occurrences, il serait intéressant de travailler, dans les mêmes conditions, avec d'autres outils d'analyse de texture comme les transformées en ondelettes ou les approches fréquentielles qui ont donné récemment de bons résultats sur la forêt. Les méthodologies proposées, aussi bien pour l'estimation des variables forestières que pour la cartographie des types de struc-

ture forestières, peuvent s'adapter et tirer avantage de ces outils complémentaires d'analyse de texture.

La méthode Random PRiF développée dans cette étude n'a été testée et comparée que sur nos jeux de données. Il apparaît nécessaire, pour démontrer son utilité et sa robustesse, de la tester sur d'autres types de forêts mais aussi sur d'autres jeux de données (issus de télédétection ou autres).

Par ailleurs, nous avons préféré dans ce travail l'utilisation de régressions linéaires qui donnent certes de bons résultats mais qui restent discutables. Des approches comme les régressions basées sur les arbres de décisions (tree-based regression), ou basées sur des approches non linéaires pourraient être envisagées et testées.

Comme nous l'avons vu précédemment, ces travaux de cartographie des types de structure forestière se basent sur l'hypothèse que les contours de parcelles sont connus. Bien entendu, une telle information n'est pas toujours disponible. Une perspective importante de ce travail serait d'évaluer le potentiel cartographique de la localisation précise de ces contours par segmentation d'image THR en utilisant des approches de classification orientées objet.

Il serait également intéressant d'approfondir l'utilisation de l'information de marge d'ensemble. La richesse de cette information pourrait être utilisée pour l'amélioration de la phase d'apprentissage via des techniques d'apprentissage actif en utilisant les instances de faibles marges qui correspondent aux instances les plus informatives pour la construction du modèle du classifieur. Elle pourrait également être utilisée dans la phase de régularisation comme post-traitement de la classification afin propager les labels des instances les plus fiables.

Enfin, l'application de notre méthodologie à l'analyse multi-dates est une perspective prioritaire, notamment pour la détection des changements. Les bons résultats obtenus semblent indiquer que la détection de changements fins de la structure forestière, cela à l'intérieur du peuplement, sera possible avec un bon niveau de précision.

## 6.5 Conclusion générale

Ce travail nous a permis de montrer le fort potentiel des images THR (et plus particulièrement des images Pléiades) pour la caractérisation du couvert forestier. Les méthodologies proposées ainsi que leurs robustesses ont été évaluées avec succès sur deux sites d'études et 4 jeux d'images THR (panchromatique et multi-spectral) acquis à différentes saisons. L'ensemble des résultats obtenus indiquent un fort potentiel pour le développement de solutions opérationnelles pour l'aide à la décision et à la gestion de la ressource forestière sur le massif de pin maritime des Landes de Gascogne.



# Références

- Barbier, N., P. Couteron, C. Proisy, Y. Malhi, and J. P. Gastellu-Etchegorry (2010). The variation of apparent crown size and canopy heterogeneity across lowland amazonian forests. *Global Ecology and Biogeography* 19, 72–84.
- Beguet, B., S. Boukir, D. Guyon, and N. Chehata (2013). Modelling-based feature selection for classification of forest structure using very high resolution multispectral imagery. *SMC'2013, IEEE Int. Conf. on Systems, Man, and Cybernetics, Manchester, UK*, 4294–4299.
- Beguet, B., N. Chehata, S. Boukir, and D. Guyon (2012). Retrieving forest structure variables from very high resolution satellite images using an automatic method. *ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences I-7*, 1–6.
- Beguet, B., N. Chehata, S. Boukir, and D. Guyon (2014). Quantification et cartographie de la structure forestière à partir de la texture des images pléiades. *Revue Française de Photogrammétrie et de télédétection (RFPT)* (208), 83–88.
- Beguet, B., N. Chehata, and S. B. D. Guyon (2014). Classification of forest structure using very high resolution pléiades image texture, igarss 2014 ieee international geoscience and remote sensing symposium, 13-18 juillet 2014, québec, canada. communication orale. *IGARSS 2014 IEEE International Geoscience and Remote Sensing Symposium. Québec, Canada.Oral*.
- Beguet, B., D. Guyon, S. Boukir, and N. Chehata (2014). Automated retrieval of forest structure variables based on multi-scale texture analysis of vhr satellite imagery. *International Journal of Photogrammetry and remote Sensing* 96, 164–178.
- Boukir, S., C. Orny, N. Chehata, D. Guyon, and J. Wigneron (2013). Détection de changements structurels sur des images satellite haute résolution. application en milieu forestier. *Traitement du signal* 6, 401–429.

- Breiman, L. (1996). Bagging predictors. *Machine Learning* 24, 123–140.
- Breiman, L. (2001, October). Random forests. *Machine Learning* 45(1), 5–32.
- Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and regression trees. *Wadsworth Inc.*
- Bruzzone, L. and S. B. Serpico (2000). A technique for feature selection in multiclass problems. *International Journal of Remote Sensing* 21(3), 549–563.
- Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological Methods and Research*, 261–304.
- Castillo, M. A., M. Ricker, and B. H. J. D. Jong (2010). Estimation of tropical forest structure from spot-5 satellite images. *International Journal of Remote Sensing* 31-10, 2767–2782.
- Chehata, N., C. Orny, S. Boukir, and D. Guyon (2011). Object-based forest change detection using high resolution satellite images. *PIA '2011, ISPRS Photogrammetric Image Analysis Conference, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Munich, Germany* 38(3), 49–54.
- Chehata, N., C. Orny, S. Boukir, and D. Guyon (2014). Object-based change detection in wind-storm damaged forest using high resolution multispectral images. *International Journal of Remote Sensing* 35(13). To appear.
- Coburn, C. A. and A. C. B. Roberts (2004). A multiscale texture analysis procedure for improved forest stand classification. *International Journal of Remote Sensing* 25, 4287–4308.
- Connors, R., M. Trivedi, and C. Harlow (1984, March). Segmentation of a high-resolution urban scene using texture operators. *Computer Vision, Graphics, and Image Processing* 25(3), 273–310.
- Couteron, P., R. Pelissier, E. A. Nicolini, and D. Paget (2005). Predicting tropical forest stand structure parameters from fourier transform of very high-resolution remotely sensed canopy images. *Journal of Applied Ecology* 42, 1121–1128.
- C.Song and C. E. Woodcock (2003). A regional forest ecosystem carbon budget model :impacts of forest age structure and landuse history. *Ecological Modelling* 164, 33–47.

- Culbert, P. D. and al (2009). The impact of phenological variation on texture measures of remotely sensed imagery. *IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing* 2(4), 299–309.
- Dash, M. and H. Liu (1997). Feature selection for classification. *Intelligent Data Analysis* 1, 131–156.
- Dell’Acqua, F. and P. Gamba (2003). Texture-based characterization of urban environments on satellite sar images. *IEEE Transactions on Geoscience and Remote Sensing* 41(1), 153–159.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–5.
- Dionne, D., J.-G. Boureau, M. Deshayes, D. Gratton, and D. Marceau (1996). Etude de la résolution spatiale optimale dans un milieu forestier de moyenne montagne. *BUL. S.F.P.T.* 141, 51–55.
- Dodge, Y. and V. Rousson (2004). *Analyse de régression appliquée*. Dunod.
- Efron, B., T. Hastie, and I. J. R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32-2, 407–499.
- Estes, L. D., P. R. Reillo, A. G. Mwangi, G. S. Okin, and H. H. Shugart (2010). Remote sensing of structural complexity indices for habitat and species distribution modeling. *Remote Sensing of Environment* 114, 792–804.
- Feng, Y., Z. Li, and T. Tokola (2010). Estimation of stand mean crown diameter from high-spatial-resolution imagery based on a geostatistical method. *International Journal of Remote Sensing* 31-2, 363–378.
- Franklin, S. E., A. J. Maudie, and M. B. Lavigne (2001). Using spatial co-occurrence texture to increase forest structure and species composition classification accuracy. *Photogrammetric Engineering & Remote Sensing* 67-7, 849–855.
- Franklin, S. E., E. Steven, M. A. Welder, A. Mike, Lavigne, and B. Mike (1996). Automated derivation of geographic window sizes for use in remote sensing image texture analysis. *Computer and Geosciences* 22(6), 665–673.
- Gaetano, R., G. Scarpa, and G. Poggi (2009). Hierarchical texture-based segmentation of multiresolution remote sensing images. *IEEE Transactions on geoscience and remote sensing* 47, 2129–2141.

- Garestier, F., P. Dubois-Fernandez, D. Guyon, and T. L. Toan (2009). Forest biophysical parameter estimation using l- and p-band polarimetric sar data. *IEEE Trans. Geosc. Remote Sens* 47(10), 3379–3388.
- Genuer, R., J. Poggi, and C. Tuleau-Malot (2010). Variable selection using random forests. *Pattern Recognition Letters* 31, 2225–2236.
- Gislason, P., J. Benediktsson, and J. Sveinsson (2006). Random Forests for land cover classification. *Pattern Recognition Letters* 27, 294–300.
- Gomez, C., M. A. Wulder, F. Montes, and J. A. Delgado (2012). Modeling forest structural parameters in the mediterranean pines of central spain using quickbird-2 imagery and classification and regression tree analysis (cart). *Remote Sensing* 4, 135–139.
- Gonzales, R. C. and R. E. Woods (2008). *Digital Image Processing*. Pearson Education Inc.
- Gromping, U. (2006). Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software* 17, 1–27.
- Guo, L. (2008). Remotely sensed data classification. Technical report, EGID Institute, University of Bordeaux 3.
- Guo, L. (2011, August). *Margin framework for ensemble classifiers. Application to remote sensing data*. Ph. D. thesis, University of Bordeaux 3.
- Guo, L., N. Chehata, C. Mallet, and S. Boukir (2011). Relevance of airborne lidar and image data for urban scene classification using random forests. *ISPRS Journal of Photogrammetry and remote Sensing* 66-1, 56–66.
- Guyon, D., I. Champion, C. Meredieu, A. Kruszewski, and T. Labbe (2005, April). Final report, work package 1\_1, ground data base. Technical report, INRA/Egypstar04/IC01, ESTEC - European Space Agency, Contract Number 18501/04.
- Guyon, D. and J. Riom (1996). Estimation de caractéristiques forestières à partir d’images à haute résolution spatiale (spot5). *BUL. S.F.P.T.* 141, 46–50.
- Guyon, I. and A. Elissee (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* 3(7-8), 1157–1182.

- Haan, C. T. (2002). *Statistical methods in Hydrology* (second ed.). Iowa State University Press, Ames, Iowa.
- Haapanen, R. and S. Tuominen (2008). Data combination and feature selection for multi-source forest inventory. *Photogrammetric Engineering & Remote Sensing* 74(7), 869–880.
- Ham, J., Y. Chen, M. Crawford, and J. Ghosh (2005). Investigation of the Random Forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* 43(3), 492–501.
- Haralick, R. M., K. Shanmugan, and I. Dinstein (1973). Texture features for image classification. *IEEE Transactions Systems, Man and Cybernetics*, 610 – 621.
- Hay, G. J., K. O. Niemann, and G. McLean (1996). An object-specific image-texture of h-resolution forest imagery. *Remote Sensing of Environment* 55, 108–122.
- Huang, C., K. Song, S. Kim, J. R. G. Townshend, P. Davis, J. G. Masek, and S. N. Goward (2008). Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment* 112(3), 970–985.
- Hyypä, J., H. Hyypä, M. Inkinen, M. Engdahl, S. Linko, and Y. H. Zhu (2000). Accuracy comparison of various remote sensing data sources in the retrieval of forest stands attributes. *Forest Ecology and Management* 128, 1-2, 109–120.
- Jolly, A. (1993). *Estimation par télédétection satellitaire de la récolte annuelle en bois dans la futaie pure de pin maritime du massif des Landes de Gascogne, Apports pour la prévision de la ressource forestière*. Ph. D. thesis, Thèse de doctorat de l'Université Paul Sabatier de Toulouse, Spécialité: Télédétection spatiale (n d'ordre: 1626), soutenue le 17 dmbre 1993, 315p.
- Jolly, A. (2014). Forestry session introduction. In *PLEIADES DAYS, 1-3 avril 2014, Toulouse, France*.
- Jolly, A., D. Guyon, and J. Riom (1993). Système d'information géographique, télédétection et politique forestière régionale. *Rev. For. Fr. XLV, N spécial Informatique et Foresterie*, 98–107.

- Jolly, A., D. Guyon, and J. Riom (1996). Utilisation des données du moyen infrarouge de landsat tm pour la mise en évidence des coupes rases sur le massif forestier landais. *International Journal of Remote Sensing*, vol.17, no.18, 3615-3645..
- Kayitakire, F., C. Hamel, and P. Defourny (2006). Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery. *Remote Sensing of Environment* 102, 390–401.
- Kerroum, M. A., A. Hammouch, and D. Aboutajdine (2010). Textural feature selection by joint mutual information based on gaussian mixture model for multispectral image classification. *Pattern Recognition Letters* 31, 1168–1174.
- Kira, K. and L. Rendell (1992). A practical approach to feature selection. In *International Conference on Machine Learning, Aberdeen, Scotland, UK*.
- Kohavi, R. and G. Jhon (1997). Wrappers for feature subset selection. *Artificial Intelligence* 97, 273–324.
- Koller, D. and M. Sahami (1996). Toward optimal feature selection. *Proceedings of the Thirteenth International Conference on Machine Learning*, 284–292.
- Kononenko, I. (1994). Estimating attributes : Analysis and extensions of relief. *European Conference on Machine Learning*, 171–182.
- Kutner, M. H., C. J. Nachtsheim, and J. Neter (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Irwin.
- LeBris, A. (2012). Extraction of vineyards out of aerial ortho-image using texture information. *ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences, Melbourne, Australia*.
- Li, S., H. Wu, D. Wan, and J. Zhu (2011). An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems* 24, 40–48.
- Lu, D. and Q. Weng (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28(5), 823–870.
- Maillard, P. (2003). Comparing texture analysis methods through classification. *Photogrammetric Engineering & Remote Sensing* 69(4), 357–367.

- Mallet, C. (2010). Analyse de donn lidar tour d'onde compl pour la classification en milieu urbain. Master's thesis, Tcom ParisTech.
- Mallows, C. L. (1973). Some comments on cp. *Technometrics* 15(4), 661-675.
- Manly, B. F. J. (1994). *Multivariate Statistical Methods : A Primer*. Chapman & Hall, London, UK.
- Mountrakis, G., J. Im, and C. Ogole (2011). Support vector machines in remote sensing : A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 66, 247–259.
- Murray, H., A. Lucieer, and R. Williams (2010). Texture-based classification of sub-antarctic vegetation communities on Heard Island. *International Journal of Applied Earth Observation and Geoinformation* 12, 138–149.
- Ni, W., C. E. Woodcock, and D. L. B. Jupp (1999). Variance in bidirectional reflectance over discontinuous plant canopies. *Remote Sensing of Environment* 69, 1–15.
- Ozdemir, I. and A. Karnieli (2011). Predicting forest structural parameters using the image texture derived from worldview-2 multispectral imagery in a dry-land forest, Israel. *International Journal of Applied Earth Observation and Geoinformation* 13, 701–710.
- Pacifici, F., M. Chini, and W. J. Emery (2009, June). A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment* 113(6), 1276–1292.
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26(1), 217–222.
- Porté, A., A. Bosc, I. Champion, and D. Loustau (2000). Estimating the foliage area of maritime pine (*Pinus pinaster* Ait.) branches and crowns with application to modelling the foliage area distribution crown. *Annals of Forest Science* 57, 73–86.
- Proisy, C., P. Couteron, and F. Fromard (2007). Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images. *Remote Sensing of Environment* 109, 379–392.

- Puig, D., M. Garcia, and J. Melendez (2010). Application-independent feature selection for texture classification. *Pattern Recognition* 43, 3282–3297.
- Puissant, A., J. Hirsch, and C. Weber (2005). The utility of texture analysis to improve per-pixel classification for high to very high spatial resolution imagery. *International Journal of Remote Sensing* 26, 733–745.
- Regniers, O., J.-P. DaCosta, G. Grenier, C. Germain, and L. Bombrun (2013). Texture based image retrieval and classification of very high resolution maritime pine forest images. *Proceedings IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.
- Robnik-Sikonja, M. and I. Kononenko (2003). Theoretical and empirical analysis of relief and rrelieff. *Machine Learning* 53(1-2), 23–69.
- Ruiz, L. A., A. Fdez-Sarria, and J. A. Recio (2004). Texture feature extraction for classification of remote-sensing data using wavelet decomposition: a comparative study. In *20th ISPRS Congress*.
- Sayes, Y., I. Inza, and P. Larranaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517.
- Schapire, R. E., T. Freund, P. Bartlett, and W. S. Lee (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. In Fisher, D. (Ed.), *Machine Learning: Proceedings of the Fourteenth International Conference. Morgan Kaufmann..*
- Schapire, R. E. and Y. Singer (1998). Improved boosting algorithms using confidence-rated predictions. *11th Annu. Conf. on Comput. Learning Theory*, 8091.
- Serpico, S. B. and L. Bruzzone (2001). A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 39(7), 1360–1367.
- Soe, W. M. and S. N. L. J. M. Tyler (2004). Wavelets for urban spatial feature discrimination: Comparisons with fractal, spatial autocorrelation, and spatial co-occurrence approaches. *Photogrammetric engineering and remote sensing* 70(7), 803–812.
- Song, C., M. B. Dickinson, L. Su, S. Zhang, and D. Yaussey (2010). Estimating average tree crown size using spatial information from ikonos and quickbird

- images: Across-sensor and across-site comparisons. *Remote Sensing of Environment* 114, 1099–1107.
- Song, C. and C. E. Woodcock (2002). The spatial manifestation of forest succession in optical imagery. the potential of multiresolution imagery. *Remote Sensing of Environment* 82, 271–284.
- Song, C. and C. E. Woodcock (2003). Estimating tree crown size from multiresolution remotely sensed imagery. *Photogrammetric Engineering & Remote Sensing*, 1263–1270.
- St-Onge, B. A. and F. Cavayas (1997). Automated forest structure mapping from high resolution imagery based on directional semivariogram estimates. *Remote Sensing of Environment* 61, 82–95.
- Thorburn, W. M. (1918). The myth of occam’s razor. *Mind* 27(107), 345–353.
- Trias-Sanz, R., G. Stamon, and J. Louchet (2008). Using colour, texture, and hierarchical segmentation for high-resolution remote-sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 63, 156–168.
- Tso, B. and P. Mather (2001). *Classification methods for remotely sensed data*. Taylor & Francis.
- Tuceryan, M. and A. Jain (1998). *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*(207-248). C. H. Chen, L. F. Pau, P. S. P. Wang.
- Tuominen, S. and A. Pekkarinen (2005). Performance of different spectral and textural aerial photograph features in multi-source forest inventory. *Remote Sensing of Environment* 94, 256–268.
- Tuttle, E. M., R. R. Jensen, V. A. Formica, and R. A. Gonsler (2006). Using remote sensing image texture to study habitat use patterns: a case study using the polymorphic white-throated sparrow (*zonotrichia albicollis*). *Global Ecology and Biogeography*, 15, 359–357.
- Van Coillie, F. M., L. P. Verbeke, and R. R. D. Wulf (2007). Feature selection by genetic algorithms in object-based classification of {IKONOS} imagery for forest mapping in flanders, belgium. *Remote Sensing of Environment* 110(4), 476 – 487.

- Vine, P., F. Bouguerzaz, and C. Puech (1996). Evaluation de l'intérêt des données spot 5 simulées pour la cartographie forestière en milieu méditerranéen. *BUL. S.F.P.T.* 141, 56–59.
- Waske, B. and J. Benediktsson (2007). Fusion of support vector machines for classification of multisensor data. *IEEE Transactions on Geoscience & Remote Sensing* 45(12), 3858–3866.
- Waske, B. and M. Braun (2009). Classifier ensembles for land cover mapping using multitemporal sar imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 64(5), 450 – 457.
- Wolter, T., P. A. Townsend, and B. R. Sturtevant (2009). Estimation of forest structural parameters using 5 and 10 meter spot-5 satellite data. *Remote Sensing of Environment* 113, 2019–2036.
- Wood, E. M., A. M. Pidgeon, V. C. Radeloff, and N. S. Keuler (2012). Image texture as a remotely sensed measure of vegetation structure. *Remote Sensing of Environment* 121, 516–526.
- Wulder, M. A., E. F. L. Drew, S. E. Franklin, and M. B. Lavigne (1998). Aerial image texture information in the estimation of northern deciduous and mixed wood forest leaf area index (lai). *Remote Sensing of Environment* 64, 64–76.
- Wunderle, A. L., S. E. Franklin, and X. G. Guo (2007). Regenerating boreal forest structure estimation using spot-5 pansharpened imagery. *International Journal of Remote Sensing* 28-19, 4351–4364.
- Yauschew-Raguenes, N. (2012). Dynamique saisonnière de la végétation forestière (arbres et sous-bois) dans le massif des landes - application de la télédétection optique au suivi des hétérogénéités à l'échelle régionale. Master's thesis, Thèse de doctorat, Université de Bordeaux 1, spécialité Physique de l'Environnement.
- Yauschew-Raguenes, N., D. Guyon, A. Kruszewski, O. Hagolle, and J. P. Wigneron (6-11 July 2008). Discrimination of the respective contributions of understory and tree canopy to the seasonal dynamics of reflectance of the maritime pine forest in the southwest of france. *IEEE International Geoscience and Remote Sensing Symposium, IGARSS'08, Boston, Massachusetts, USA Vol. 3, III-808 - III811*.
- Zhang, Y. (2010). An improved iamb algorithm for markov blanket discovery. *Journal of Computers* 5(11), 1755–1761.

Zhu, M. (2008). Kernels and ensembles: Perspectives on statistical learning. *The American Statistician* 62(2), 97–109.



# Annexes



## Annexe A

Protocole de mesures au sol de variables de structure forestière sur un échantillon de peuplements de pin maritime sur le site Tagon-Marcheprime.

# **Protocole de mesures au sol de variables de structure forestière sur un échantillon de peuplements de pin maritime sur le site Tagon-Marcheprime**

**Date :** 24 Février 2012, dernière révision : 04 juin 2012

**Auteur :** B. BEGUET, D. GUYON

## **Proposition de protocole suite à la réunion du 23 janvier 2012**

Participants réunion :

UR EPHYSE : B. BEGUET, D. GUYON, C. MOISY, A. KRUSZEWSKI

UE Pierroton : F. BERNIER, B. ISSENHUT, P. PASTUSZKA.

## **Mots-clés :**

Dendrométrie, pin maritime, protocole, THR

## **Résumé :**

Protocole de mesure au sol de variables de structure forestière sur un échantillon de peuplements de pin maritime sur le site Tagon-Marcheprime en vue de la création d'un jeu de données d'apprentissage pour la reconnaissance des structures forestières à partir d'images à Très Haute Résolution (THR).

## **Objectifs :**

Réalisation d'un inventaire forestier sur le bassin versant du Tagon.

Appui aux projets ICLEF-THR et STELLA (financement CNES/TOSCA et Région Aquitaine) sur la caractérisation de la structure forestière et de ses changements brutaux (tempête) par télédétection à très haute résolution spatiale (0.50-2.50m).

Données nécessaires à la calibration et validation des méthodes et algorithmes de traitement d'image développés durant la thèse de Benoit Beguet.

Mesures de variables de structure forestière sur un échantillon représentatif de la diversité de la structure forestière de la zone.

Les principales variables d'intérêt sont la densité, la circonférence des troncs à 1.3m, la hauteur dominante, le diamètre des couronnes et l'âge.

Besoin d'un nombre maximal d'observations (statistiques).

**Période :** juin-septembre 2012

## 1- Localisation de la zone d'étude :

Le site d'étude se situe entre la ville de Marcheprime et la ville de Biganos, le long du bassin versant du Tagon.

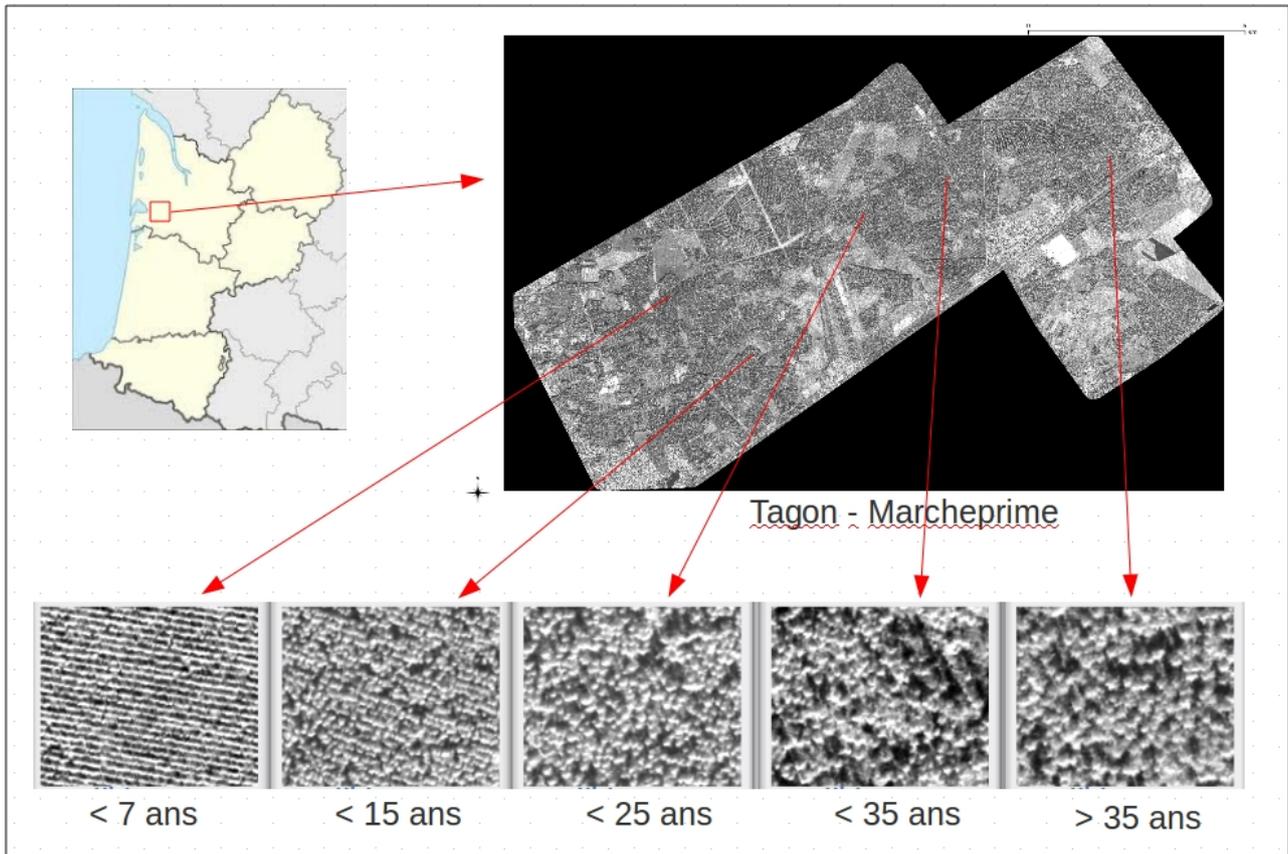


Figure 1 : Localisation de la zone d'étude, exemple de textures d'images pour des classes d'âges, image Geoeye feb 2009, (0.5m).

→ système de projection et coordonnées géographiques : WGS84 UTM 30 Nord

## 2- Parcelles échantillonnées:

- Le choix des parcelles d'intérêt se fait par photo-interprétation sur la base des images satellite Geoeye 16 février 2009 (résolution spatiale : 0.5m) et photographies aériennes 26 février 2009 (résolution spatiale : 0.15m). La délimitation des parcelles est réalisée à partir des mêmes données.

6 classes de textures d'images sont définies visuellement (elles correspondent à 6 classes de structures des peuplements qui résultent des pratiques sylvicoles et de la croissance des arbres).

21 parcelles sont sélectionnées pour chacune des classes (lorsque c'est possible) de manière à balayer un maximum de variabilité intra-classe (+10 parcelles supplémentaires par classes au cas où).

Pas de mesures dendrométriques pour la classe 1 : observation visuelle (état de surface (labour type de végétation), reboisement oui/non ancienneté).

Classe 1	Coupes rases ou reboisements récents (labour, lignes de semis, plantation ou de préparation de semis visibles), age du peuplement $\leq$ 3ans (saison 2012 comprise)
Classe 2	Rangs / arbres et ombres discernables / couvert non fermé sur l'inter rang / DiamC < 1m
Classe 3	Rangs serrés / arbres confondus / DiamC $\sim$ 1.5-2.5m (inférieurs à 3m)
Classe 4	Rangs moins visibles / arbres discernables / densité généralement forte / DiamC $\sim$ 3-4m (inférieurs à 4.5m)
Classe 5	Rangs rarement visibles / arbres discernables / densité généralement moyenne / DiamC $\sim$ 4.5-6m (inférieurs à 6.5m)
Classe 6	Rangs rarement visibles / arbres discernables / densité généralement faible / DiamC $\sim$ 6.5m et plus

Tableau 1 : Descriptif des classes de structures forestières, DiamC : diamètre des couronnes estimés visuellement sur l'image Geoeye de février 2009).

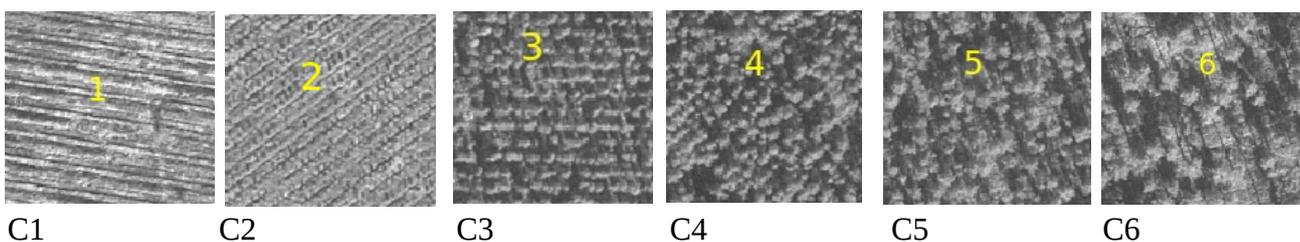


Figure 2: Exemples d'imagettes de peuplements correspondant aux classes définies (image Geoeye 16 février 2009).

Entre 2009 (image Geoeye) et 2012 (terrain + images Pléiades qui doivent être prises) un « glissement » est attendu (au moins pour les classes 1 et 2). Il faudra donc ajouter des parcelles correspondant à la classe 1 en 2012.

- Les parcelles sont échantillonnées dans trois zones différentes afin de prendre en compte le maximum de diversité dans les pratiques sylvicoles, la fertilité ainsi que la taille et la forme des peuplements.

- Sera établi un fichier de type vecteur (dendro\_tagon\_2009.shp) comprenant les limites des parcelles, un identifiant de parcelle ainsi que la classe de structure correspondante.

### 3- Principe d' échantillonnage dans chaque parcelle :

Pour chaque parcelle, trois placettes de mesures sont échantillonnées.

Le centre d'une placette doit être au moins à 40m des bords de la parcelle.

Taille d'une placette :

- un carré de 20m de coté si les lignes de pins sont bien visibles
- sinon disque de de 400m<sup>2</sup> (rayon = 11.28m)
- si la placette contient moins de 10 arbres vivants: la placette doit être élargie de manière à contenir 10 arbres et la distance entre le centre de la placette et l'arbre le plus éloigné (le 10ième) doit être mesurée.

Si la parcelle est trop petite (largeur de la parcelle < 160m), le nombre de placettes est réduit en conséquence. La taille minimale d'une parcelle est de 80\*80m.

Une distance minimale de 80m entre centres des placettes doit être respectée (60m entre les bords des placettes).

Un relevé Gps doit être réalisé au centre de la placette (ou proche du centre suivant ce qui est possible). Système de projection et coordonnées géographiques : WGS-UTM30N

La figure ci-dessous représente la configuration souhaitée :

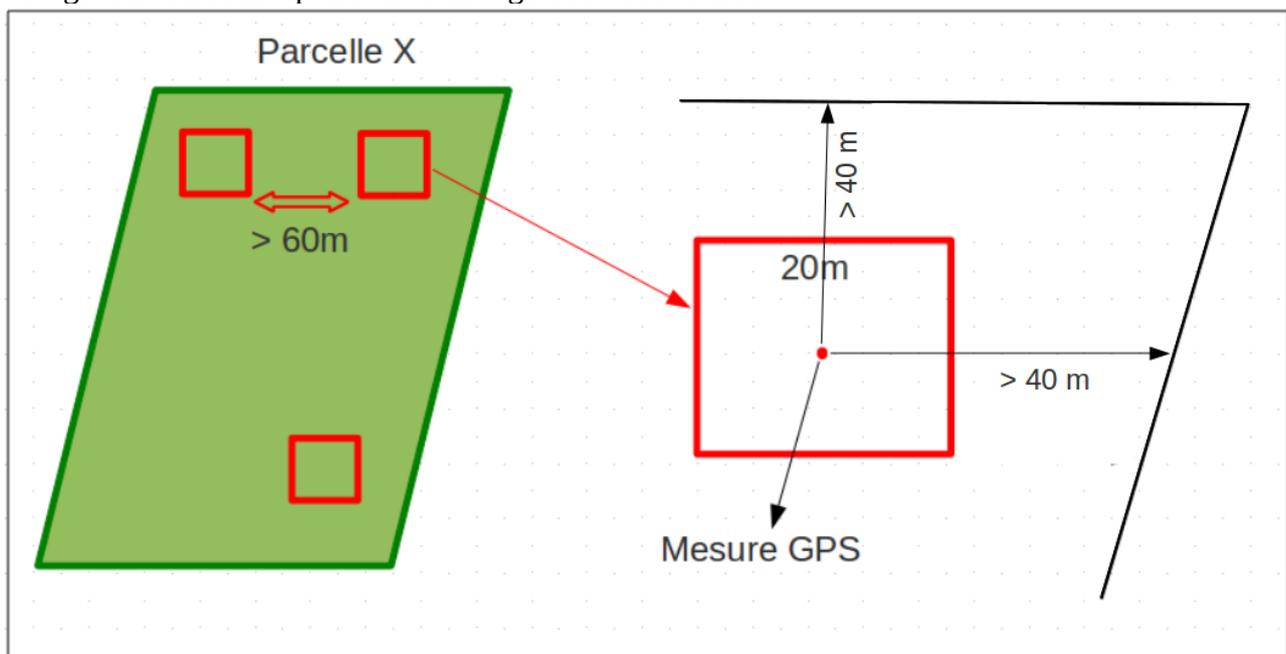


Figure 3 : Représentation schématique du dispositif de mesure

#### 4- Déroulement de l'échantillonnage :

→ Emboitement :

- 7 parcelles par zones pour chaque classe
- A chaque sortie, l'objectif est d'avoir un équilibre dans la distribution des classes de structures dans les zones.

Codage du nom des parcelles et placettes :

<i>N° Zone</i>	<i>N° Classe</i>	<i>N° Parcelle</i>	<i>N°ordre placette</i>
1 à 3	1 à 6	01 à 99	1 à 3

Dans un premier temps, l'objectif est de mesurer une placette par parcelle afin d'avoir au moins une placette pour chacune des parcelles considérées (105 placettes n°1 pour les classes supérieures à 1).

L'emplacement de ces placettes n°1 est choisi par l'UE.

Une fois ces premières placettes mesurées, les placettes n°2 et 3 seront mesurées.

#### 5- Documents de travail fournis à l'UE :

- le présent protocole
- fichiers SIG (contours et identifiants des parcelles) : parcellaire initial + parcellaire complémentaire
- cartes papier :
  - fond IGN 1/25000 localisation des zones et parcelles (avec identifiants à 4 numéros)
  - zoom par zones sur images satellite la plus récente possible.

#### 6- Variables mesurées:

Ces mesures concernent les classes 2 à 6 (age > 3ans, saison 2012 comprise).

Classe 1 : observation visuelle uniquement. Si cette parcelle appartient à la classe 2 au moment des mesures, la mesurer comme telle.

##### 1. Estimation de la densité :

Sur chaque placette, comptage du nombre d'arbres vivants/morts, et :

- Si les lignes d'arbres sont visibles : dénombrement des arbres sur chaque portion de ligne et mesure de l'espacement moyen des lignes.
- Sinon :
  - placette circulaire =400m<sup>2</sup> : comptage du nombre d'arbres dans un disque de 11.28 m de rayon centré sur le point de référence.

- placette >400m<sup>2</sup> de 10 arbres vivants: mesure de la distance entre l'arbre vivant le plus éloigné du centre et le centre (Ne pas oublier de dénombrer les morts à distance du centre < à celle du 10ième vivant).

## 2. Circonférence des troncs :

Sur chaque placette, les 10 arbres les plus proches du centre sont choisis pour réaliser cette mesure. La mesure se fait à une hauteur de 130 cm.

## 3. Hauteur :

5 arbres vivants sont mesurés pour la placette n°1 uniquement (les plus proches du centre).

## 4. Diamètre de couronne:

Sur chaque placette, 5 arbres vivants les plus proches du centre de la placette (les mêmes que pour la mesure de hauteur).

Mode opératoire :

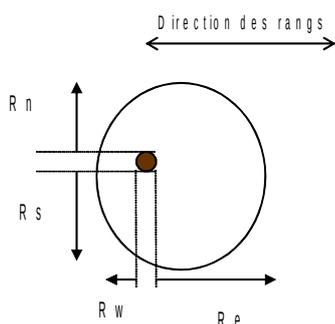
Kronenspiegel (ou perche 9m lorsque les arbres sont petits)

4 rayons, 2 perpendiculaires aux rangs, 2 parallèles (si les rangs de pin ne sont pas visibles, prendre une direction de référence commode).

Le diamètre de la surface projetée verticalement à 1.3m par la couronne est mesuré dans 2 directions : celle des rangs et celle perpendiculaire. Pour faciliter la mesure, c'est le centre de la tige à 1.30m, et non pas la projection verticale du sommet de l'arbre, qui est pris comme centre de rayonnement. On mesure du tronc (depuis sa surface extérieure) à l'extrémité de la branche verte la plus longue dans le cadran correspond à la direction donnée.

Dans chaque direction, deux « rayons » sont mesurés :  $R_w$  et  $R_e$ ,  $R_n$  et  $R_s$ . Le rayon peut être noté négatif, dans le cas très rare d'une forte excentricité de la couronne. Généralement le tronc de l'arbre est inclus dans le projection verticale de la couronne.

L'axe « Est-Ouest » correspond à la direction des rangs.



Le diamètre de la couronne  $D$  d'un arbre peut être estimé ainsi :

$$D = \frac{1}{2} \sum_{d \in \{N, E, W, S\}} (R_d) + \frac{c130}{4\pi} \sum_{d \in \{N, E, W, S\}} (W_d)$$

où  $R_d$  = rayon du houppier mesuré dans la direction  $d$ ,  $c130$  est la circonférence de la tige à 1.30m,  $W_d = 1$  si  $R_d \geq 0$ , sinon  $W_d = -1$

## 5. Age du peuplement :

Uniquement sur placette n°1.

L'âge du peuplement est estimé en comptant les verticilles sur les 5 arbres choisis pour mesurer le diamètre des couronnes, ceci lorsque le nombre de verticilles est dénombrable.

Lorsque les verticilles ne sont pas dénombrables, fournir une fourchette d'estimation de l'âge (exemple : 30-40 ans)

## 7. Observations supplémentaires et modalités de notation :

- Date de la mesure
- Stade du peuplement : coupe rase /  $\leq 3$ ans /  $> 3$  ans
- Type de reboisement : plantation / semis ligne / semis bande / non identifié
- Rejet de la parcelle : obstacle (ex :grillage) / impénétrable (ex : broussailles) / mortalité trop importante
- Forme de la placette : carrée / circulaire / circulaire  $>400m^2$
- Type de sous-bois :

	Taux de recouvrement				
	0	25	50	75	100
Molinie et autres graminées					
Fougères					
Bruyères					
Ajoncs					
Pins (régénération)					
Arbustes feuillus					
Sol nu					

- Débroussaillage : avant 2009 / depuis 2009
- Éclaircie : avant 2009 / depuis 2009
- Tempête 2009: peuplement apparemment non touché / chablis / étêtés / penchés / nettoyé
- Mortalité sur pied : scolytes / chenilles PP / autres

## Annexe B

### Analyse comparative des méthodes de sélection d'attributs : tableaux de résultats complémentaires

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.859	8.688	0.850	1.000
	LARS stepwise	0.997	0.192	0.126	1.136
	Random PriF	0.992	0.461	0.196	2.106
Sp	PCA	0.927	4.367	0.603	1.000
	LARS stepwise	0.994	0.535	0.211	1.025
	Random PriF	0.996	0.266	0.149	2.282
Ht	PCA	0.910	57.933	2.197	1.000
	LARS stepwise	0.992	5.114	0.652	1.070
	Random PriF	0.994	2.253	0.433	1.891
Nah	PCA	0.860	437362.615	190.910	1.000
	LARS stepwise	0.969	100160.933	91.360	1.315
	Random PriF	0.991	22609.138	43.406	2.242
Dbh	PCA	0.883	0.028	0.049	1.000
	LARS stepwise	0.986	0.003	0.016	1.322
	Random PriF	0.993	0.001	0.009	1.401

Tableau B.1 : Feature selection methods comparison : performance of the best 3 features model using the MS dataset of october image on Nezer site.

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.686	206.650	1.364	1.000
	LARS stepwise	0.784	142.154	1.133	1.289
	Random PriF	0.799	138.673	1.117	2.163
Sp	PCA	0.720	120.216	1.040	1.000
	LARS stepwise	0.824	75.808	0.826	1.274
	Random PriF	0.822	76.114	0.828	2.339
Ht	PCA	0.724	1774.745	3.998	1.000
	LARS stepwise	0.446834	1084.863	3.126	1.333
	Random PriF	0.804	1253.808	3.360	1.899
Nah	PCA	0.474	48495207.092	660.979	1.000
	LARS stepwise	0.619	45006219.333	636.758	1.733
	Random PriF	0.616	39528184.643	596.749	1.766
Dbh	PCA	0.712	723.519	0.080	1.000
	LARS stepwise	0.815	471.768	0.065	1.896
	Random PriF	0.804	499.595	0.067	2.109

Tableau B.2 : Feature selection methods comparison : performance of the best 5 features model using the MS dataset of august image on Tagon site.

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.915	4.655	0.622	1.000
	LARS stepwise	0.992	0.356	0.172	1.310
	Random PriF	0.994	0.313	0.161	2.025
Sp	PCA	0.870	7.944	0.813	1.000
	LARS stepwise	0.989	0.640	0.230	1.270
	Random PriF	0.996	0.139	0.107	2.250
Ht	PCA	0.876	77.011	2.533	1.000
	LARS stepwise	0.972	15.975	1.153	1.327
	Random PriF	0.993	2.760	0.479	2.704
Nah	PCA	0.745	713641.815	243.865	1.000
	LARS stepwise	0.995	19946.779	40.770	1.018
	Random PriF	0.992	13180.532	33.141	2.159
Dbh	PCA	0.879	0.027	0.047	1.000
	LARS stepwise	0.997	0.004	0.019	1.056
	Random PriF	0.993	0.001	0.010	1.376

Tableau B.3 : Feature selection methods comparison : performance of the best 3 features model using the Pan dataset of october image on Nezer site.

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.534	331.11.4460	1.727	1.000
	LARS stepwise	0.752	180.304	1.274	1.357
	Random PriF	0.800	129.015	1.078	2.293
Sp	PCA	0.548	214.820	1.391	1.000
	LARS stepwise	0.765	116.219	1.023	2.001
	Random PriF	0.808	84.263	0.871	1.860
Ht	PCA	0.556	2904.690	5.115	1.000
	LARS stepwise	0.837	1144.369	3.210	2.052
	Random PriF	0.851	1032.486	3.049	1.666
Nah	PCA	0.225	66281517.350	772.742	1.000
	LARS stepwise	0.523	51588534.708	681.734	1.408
	Random PriF	0.500	46197877.372	645.133	1.748
Dbh	PCA	0.541	1231.149	0.105	1.000
	LARS stepwise	0.818	475.571	0.065	3.817
	Random PriF	0.823	461.019	0.064	2.132

Tableau B.4 : Feature selection methods comparison : performance of the best 5 features model using the Pan dataset of august image on Tagon site.

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.917	5.458	0.674	1.000
	LARS stepwise	0.995	0.190	0.125	1.288
	Random PriF	0.996	0.168	0.118	2.421
Sp	PCA	0.906	6.149	0.715	1.000
	LARS stepwise	0.994	0.535	0.211	1.025
	Random PriF	0.996	0.233	0.139	2.734
Ht	PCA	0.903	66.034	2.345	1.000
	LARS stepwise	0.992	5.114	0.652	1.070
	Random PriF	0.994	2.791	0.482	2.519
Nah	PCA	0.871	464567.648	196.758	1.000
	LARS stepwise	0.998	8154.343	26.067	1.287
	Random PriF	0.997	5582.335	21.568	3.404
Dbh	PCA	0.892	0.029	0.049	1.000
	LARS stepwise	0.987	0.003	0.016	1.226
	Random PriF	0.993	0.001	0.009	1.270

Tableau B.5 : Feature selection methods comparison : performance of the best 3 features model using the MS-Pan dataset of october image on Nezer site.

Forest variable	Feature selection method	multiple $R^2$	PRESS	LOO RMSE	Max(VIF)
Cd	PCA	0.658	224.809	1.423	1.000
	LARS stepwise	0.797	139.073	1.119	1.098
	Random PriF	0.807	133.708	1.097	2.261
Sp	PCA	0.696	130.387	1.083	1.000
	LARS stepwise	0.837	70.821	0.798	1.154
	Random PriF	0.819	77.910	0.837	2.300
Ht	PCA	0.695	1940.319	4.180	1.000
	LARS stepwise	0.849	987.795	2.983	1.517
	Random PriF	0.831	1084.890	3.126	1.778
Nah	PCA	0.457	50573696.810	674.995	1.000
	LARS stepwise	0.633	44959535.726	636.428	1.409
	Random PriF	0.625	38113219.156	585.971	1.708
Dbh	PCA	0.684	794.497	0.084	1.000
	LARS stepwise	0.828	450.154	0.063	1.070
	Random PriF	0.808	481.756	0.065	2.070

Tableau B.6 : Feature selection methods comparison : performance of the best 5 features model using the MS-Pan dataset of august image on Tagon site.

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.968	0.783
	0.018	0.173
	0.012	0.043
Sp	0.959	0.576
	0.035	0.353
	0.004	0.069
Ht	0.975	0.555
	0.018	0.344
	0.006	0.099

*Tableau B.7 : Comparison of subsets balance between LARS stepwise and Random PRiF, Nezer Site, MS dataset*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.916	0.589
	0.071	0.275
	0.012	0.135
Sp	0.901	0.806
	0.047	0.107
	0.050	0.083
Ht	0.842	0.480
	0.089	0.460
	0.062	0.058

*Tableau B.8 : Comparison of subsets balance between LARS stepwise and Random PRiF, Nezer Site, Pan dataset*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.921	0.752
	0.011	0.272
	0.066	0.220
Sp	0.959	0.445
	0.035	0.442
	0.004	0.111
Ht	0.975	0.372
	0.018	0.326
	0.006	0.300

*Tableau B.9 : Variable importance in subsets provided by LARS stepwise and Random PRiF feature selection methods, Nezer site, MS-PAN dataset.*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.766	0.680
	0.095	0.143
	0.079	0.047
	0.052	0.046
	0.006	0.08
Sp	0.625	0.689
	0.212	0.150
	0.126	0.125
	0.017	0.023
	0.017	0.010
Ht	0.709	0.564
	0.144	0.240
	0.072	0.104
	0.040	0.045
	0.031	0.045

*Tableau B.10 : Variable importance in subsets provided by LARS stepwise and Random PRiF feature selection methods, Tagon site (June), MS dataset.*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.761	0.788
	0.126	0.141
	0.069	0.040
	0.037	0.024
	0.005	0.005
Sp	0.811	0.613
	0.124	0.180
	0.051	0.148
	0.005	0.029
	0.006	0.028
Ht	0.674	0.599
	0.154	0.234
	0.142	0.074
	0.019	0.061
	0.008	0.029

*Tableau B.11 : Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). MS dataset*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.563	0.719
	0.277	0.134
	0.099	0.093
	0.039	0.033
	0.019	0.018
Sp	0.462	0.500
	0.254	0.275
	0.211	0.150
	0.053	0.071
	0.017	0.002
Ht	0.410	0.685
	0.328	0.193
	0.124	0.076
	0.068	0.037
	0.068	0.007

*Tableau B.12 : Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). Pan dataset*

Forest variable	Texture variable importance	
	LARS Stepwise	Random PRiF
Cd	0.780	0.783
	0.089	0.139
	0.050	0.034
	0.044	0.023
	0.034	0.021
Sp	0.845	0.765
	0.091	0.184
	0.010	0.032
	0.045	0.013
	0.006	0.008
Ht	0.549	0.575
	0.275	0.166
	0.150	0.136
	0.013	0.117
	0.011	0.005

*Tableau B.13 : Comparison of subsets balance between LARS stepwise and Random PRiF, Tagon Site (August). MS-Pan dataset*

## Annexe C

Retrieving forest structure variables from very high resolution satellite images using an automatic method.

## RETRIEVING FOREST STRUCTURE VARIABLES FROM VERY HIGH RESOLUTION SATELLITE IMAGES USING AN AUTOMATIC METHOD

B.Beguet<sup>1,2</sup>, N.Chehata<sup>1,3</sup>, S. Boukir<sup>1</sup> and D. Guyon<sup>2</sup>

<sup>1</sup>EA 4592 G&E Laboratory, University of Bordeaux, 1 Allée F.Daguin, 33607 Pessac Cedex, France

<sup>2</sup>INRA, UR1263 EPHYSE, 33140 Villenave d'Ornon, France

<sup>3</sup>LISAH UMR 144, IRD, El Menzah 4, Tunis, Tunisia

<sup>1</sup>name.surname@egid.u-bordeaux3.fr, <sup>2</sup>guyon@bordeaux.inra.fr

Commission VII/1

**KEY WORDS:** Forestry, Modelling, Texture, Multi-resolution, Quickbird

### ABSTRACT:

The main goal of this study is to define a method to describe the forest structure of maritime pine stands from Very High Resolution satellite imagery. The emphasis is placed on the automatization of the process to identify the most relevant image features, exploiting both spectral and spatial information. Our approach is based on linear regressions between the forest structure variables to be estimated and various spectral and Haralick's texture features (derived from Grey Level Co-occurrence Matrix). The main drawback of this well-known texture representation is the underlying parameters (window size, displacement length, orientation and quantification level) which are extremely difficult to set due to the spatial complexity of forest structure. To tackle this major issue, probably the main cause of poor texture analysis in practice, we propose an automatic feature selection process whose originality lies on the use of image test frames of adequate forest samples whose forest structure variables were measured at ground. This method, inspired by camera calibration protocols, selects the best image features via statistical modelling, exploring a wide range of parameter values. Hence, just a few samples are required to build up the test frames but allow a fast assessment of thousands of descriptors, given the large number of tested combinations of parameters values. This method was developed and tested on Quickbird panchromatic and multispectral images. It has been successfully applied to the modelling of 7 typical forest structure variables (age, tree height, crown diameter, diameter at breast height, basal area, density and tree spacing). The coefficient of correlation,  $R^2$ , of the best single models for 6 of the forest variables of interest, estimated from the test frames, ranges from 0.89 to 0.97. Only the basal area was weakly correlated to the considered image features (0.64). To improve the results, combinations of panchromatic and or multi-spectral features were tested using multiple linear regressions. As collinearity is a very perturbing problem in multi-linear regression, this issue is carefully addressed. Different variables subset selection methods are tested. A new stepwise method, derived from LARS (Least Angular Regression), turned out the most convincing, significantly improving the quality of estimation for all the forest structure variables ( $R^2 > 0.98$ ). Validation is done through stand ages retrieval along the whole site. The best estimation results are obtained from subsets combining multi-spectral and panchromatic features, with various values of window size, highlighting the potential of a multi-scale approach for retrieving forest structure variables from VHR satellite images.

### 1 INTRODUCTION

Many studies have focused on estimating forest parameters from remote sensing data since the early years of satellite imagery. The accuracy in retrieving forest stand variables such as stem volume, tree height and basal area from different image sources depends on their spatial resolution (Hyypä et al., 2000). While SPOT or Landsat TM data give low correlation performance, aerial photographs provide highly significant information. Using SPOT images, (Wunderle et al., 2007, Wolter et al., 2009, Castillo et al., 2010) retrieve some forest stand attributes, exploiting image texture, with a good accuracy. Over the last decade, a growing number of Very High Resolution (VHR) remote sensing data from various spatial sensors has become available. The spatial resolution of the provided images is comparable to the one of aerial photographs. Some recent studies (Kayitakire et al., 2006, Feng et al., 2010, Proisy et al., 2007, Ozdemir and Karnieli, 2011, Song et al., 2010) have shown the potential of VHR imagery for forest inventory applications thanks to the strong relationship between forest spatial structure and image texture at stand level. In this paper, we aim to fully exploit the potential of texture features, extracted from VHR satellite images, in describing typical forest variables, with a particular emphasis on automatic parameter tuning, one of the major issues in texture analysis. VHR imagery provides a meaningful textural information. How-

ever, no unique optimum method exists to properly describe and use this complex information. Some relevant methods are based on Fourier spectra calculation (Barbier et al., 2010), geometric (M.Tuceryan, 1998) or morphologic (Huang et al., 2008) approaches. Texture analysis of VHR satellite images generally applied to forest inventory can be divided into two main approaches: variogram (Guyon and Riou, 1996, Song et al., 2010) and GLCM (Grey Level Co-occurrence Matrix) (Franklin et al., 2001, Kayitakire et al., 2006, Murray et al., 2010, Coburn and Roberts, 2004, Wunderle et al., 2007, Wolter et al., 2009, Castillo et al., 2010, Ozdemir and Karnieli, 2011). A variogram characterizes image spatial properties with a plot of semi-variance as a function of distance between scene components. Its main limit is the necessity of selecting a good non-linear model to fit the variogram before extracting indicators such as sill, range and nugget. GLCM are second order statistics constructed by 2D histograms. While first order statistics are generally calculated for comparison purposes, texture features derived from GLCM are the most used in recent remote sensing literature. It provides good performances for forest parameters estimations and its implementation remains less complex than other texture representations like wavelets or Gabor filters. The main drawback of this well-known texture representation is the underlying parameters (window size, displacement length, orientation and quantification level) which are extremely difficult to set due to the spatial complexity of the forest struc-

ture. To tackle this major issue, probably the main cause of poor texture analysis in practice, an automatic feature selection process that explores a wide range of texture parameter values is a worthwhile solution to investigate.

The objective of the present study is to provide a fully automatic method to retrieve the most correlated texture features with the following forest structure variables: crown diameter, density and tree spacing which induce less or more texture in the image depending on the image spatial resolution. Complementary forest structure variables are also considered as tree height, diameter at breast height and basal area, as well as stand age which can be considered as a synthetic forest variable. The height of top of tree also contributes to image texture, in particular since it impacts on the length of its shadows according to the solar elevation or to the view zenith angle of the sensor. Texture analysis is based on GCLM with an automatic selection process of the features and their parameterization.

To achieve optimal results, some combinations of panchromatic and/or multi-spectral features were tested using multiple linear regressions between ground measured variables and image features. As collinearity is a very perturbing problem in multi-linear regression, this issue is carefully addressed through a new variable subset selection algorithm that has never been involved in remote sensing forestry : LARS (Least Angular Regression) (Efron et al., 2004), and outperformed other existing variable subset selection techniques. The developed method was assessed on maritime pine stands covering a large forest structure diversity from Quickbird images.

## 2 MATERIAL

### 2.1 Study site

The Nezer site is located in a large forest of maritime pine (*Pinus pinaster Ait.*) in southwestern France. Its surface is nearly flat. It is composed of maritime pine even-aged stands which are intensively managed. Their size is variable from approximately 4 to 50 ha. They are often circled by firebreak clearings or roads. The pine trees are planted in rows 4 m apart, which are oriented West-East in most of stands.

### 2.2 Field data

Measurements of forest structural variables were made on 12 stands at the end of year 2003. The trees sampled in each stand were located within a 80m×80m square plot representative of the forest structure in the 200m×200m area that encloses it. Crown diameter (Cd), stand density (Nah), tree Height (Ht), Diameter at breast height (Dbh), and Basal area (Ba) were measured. Tree Spacing (Sp) is estimated as an empirical function of density, this allows a linearization of density variations.

$$Sp(m) = \sqrt{\frac{20000}{\sqrt{3} \text{density}(tree/ha)}} \quad (1)$$

Age of the sampled stands is ranged from 13 to 51 years. The variation range of the forest variables are presented in Table 1. As expected, due to allometric constraints, a strong correlation is observed between age, Cd, Dbh, Ht, Nah and Sp. Only Ba, which strongly depends on the changes in density and Dbh due to thinning, is not as correlated to other structural variables. The exact age is known for all stands of the site from forest manager's archives. Only 184 stands, with available remote sensing data, are involved in this study.

### 2.3 Remote sensing data

Quickbird is among the satellite sensors with the highest spatial resolution and its spectral characteristics are suitable for forest

	Cd (m)	Nah (tree/ha)	Sp (m)	Ht (m)	Dbh (m)	Ba (m <sup>2</sup> /ha)	Age (year)
Min	2.9	189	3.0	9.2	0.15	20.2	13
Max	7.8	1253	7.8	24.6	0.46	33.2	51

Table 1: Range of variations of forest variables over 12 sampled stands

applications. This sensor has a spatial resolution ranging from 0.61 to 0.70 m in the panchromatic channel (Pan) and from 2.44 to 2.88 in the multi-spectral channels (MS), depending on the view angle, and a radiometric resolution of 11 bits. The image, acquired on the 6th of October 2003, was delivered as a sensor corrected standard product of an excellent quality, with no clouds and a view angle of 19°. All channels (MS and Pan) were well co-registered (registration error < 1 pixel). As the image was acquired in October, a period where under-tree green biomass amount is rather low, the spatial variations in understory vegetation reflectance were probably low.

## 3 METHODS

### 3.1 Features

We considered first order texture features (mean and variance) and second order GLCM texture features to describe the spatial relationship between a pixel and its neighborhood. (Haralick et al., 1973) defined 14 texture features derived from GLCM. Some of them are considered as particularly relevant for image analysis on forest applications (Franklin et al., 2001, Coburn and Roberts, 2004, Kayitakire et al., 2006, Murray et al., 2010, Wunderle et al., 2007, Ozdemir and Karnieli, 2011, Castillo et al., 2010). These studies generally involved one spatial resolution, either panchromatic or multi-spectral. To our knowledge, only Wolter (Wolter et al., 2009) has combined features at both resolutions, but from variogram and not GLCM as in the present work. All texture measures were calculated on both Panchromatic and Multi-spectral bands. We used the four available spectral bands (Blue (B), Green (G), Red (R), Near Infra Red (NIR)) and the well-known NDVI (Normalized Difference Vegetation Index). We considered the eight most used Haralick's features in remote sensing forestry (Wunderle et al., 2007, Ozdemir and Karnieli, 2011, Castillo et al., 2010) :

$$\text{Energy (energy)} = f_1 = \sum_{i,j} g(i,j)^2$$

$$\text{Entropy (entropy)} = f_2 = - \sum_{i,j} g(i,j) \log_2 g(i,j), \text{ or } 0 \text{ if } g(i,j) = 0$$

$$\text{Correlation (corr)} = f_3 = \sum_{i,j} \frac{(i-\mu)(j-\mu)g(i,j)}{\sigma^2}$$

$$\text{Inverse Difference (invdif)} = f_4 = \sum_{i,j} \frac{1}{1+(i-j)^2} g(i,j)$$

$$\text{Inertia (inertia)} = f_5 = \sum_{i,j} (i-j)^2 g(i,j) \text{ (sometimes called contrast)}$$

$$\text{Cluster Shade (clushade)} = f_6 = \sum_{i,j} ((i-\mu) + (j-\mu))^3 g(i,j)$$

$$\text{Cluster Prominence (cluspro)} = f_7 = \sum_{i,j} ((i-\mu) + (j-\mu))^4 g(i,j)$$

$$\text{Haralick's Correlation (harcorr)} = f_8 = \frac{\sum_{i,j} (i,j)g(i,j) - \mu_t^2}{\sigma_t^2} \text{ where}$$

$\mu_t$  and  $\sigma_t$  are the mean and standard deviation respectively of the row (or column, due to symmetry) sums. Above,  $\mu = \sum_{i,j} i \cdot g(i,j) = \sum_{i,j} j \cdot g(i,j)$  is the weighted pixel average, and  $\sigma = \sum_{i,j} (i-\mu)^2 \cdot g(i,j) = \sum_{i,j} (j-\mu)^2 \cdot g(i,j)$  (due to matrix symmetry) is the weighted pixel variance.

Besides, we added another feature, pantex, which represents the min value of the contrast in 8 directions (Pesari et al., 2008). This provides an anisotropic feature. GLCM parameters are: radius of the moving window  $r$  (*window size* =  $2r+1$ ), displacement  $d$ , orientation  $o$  and quantification number  $nbbin$ . We chose to limit the range of Haralick parameters to experiment to achieve a trade-off between accuracy and efficiency. We defined different ranges with respect to spatial resolutions (Table 2). As for the quantification level, since only forest pixels (image samples) are

considered, setting nbbin=8, as traditionally used in literature, appears to be sufficient. As there is no way of knowing *a priori* the

Data	$r$ (pixel)	$d$ (pixel)	$\theta$ (degree)
Pan	5-25 step: 5	1-10 step: 1	0-135 step: 45
MS	3-12 step: 3	1-4 step: 1	0-135 step: 45

Table 2: GLCM parameter range values

appropriate GLCM parameters to use for our data, we chose to explore a wide range of parameter values. The first order features were calculated using the same window radius values.

### 3.2 Automatic feature selection

**Test frame :** We propose an automatic feature selection process whose originality lies on the use of test frames of adequate forest samples where the structure variables were measured from ground. This method, inspired by camera calibration protocols, selects the best image features via statistical modelling, exploring a wide range of parameter values. Hence, just a few samples are required to build up the test frames but allow a fast assessment of thousands of descriptors, given the large number of tested combinations of parameters values. Compared to more traditional supervised methods, which are very costly in terms of field surveys and time-consuming, our method requires a slight supervision. Furthermore, in the context of texture analysis, which involves high computational costs, the use of test frames appears as an appealing way to explore a wide range of texture parameters which would be impractical in case of large image samples. The forest test frame is constructed with 12 small image samples corresponding to the field measurements (cf. Figure 1). Square plot areas are centered on the field measurement area, with a width of 200 pixels for Pan and 50 pixels for MS (around 120 meters on the ground). Maximum and minimum values of the whole test frame are used to construct the GLCM's quantification levels (this guarantees that only *forest* pixels will be used). For each square plot, the average (M) and the variance (V) of all features were saved for statistical analysis.

Two approaches were tested for statistical modelling; linear regression where a forest variable is described by a single texture feature and thus exploiting only one spatial resolution. The second approach is a multi-scale method based on multiple regressions using several Pan and MS features to estimate a given forest variable.

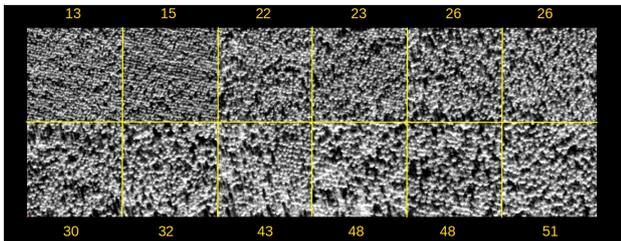


Figure 1: Panchromatic test frame, sorted by increasing age

**Single variable linear regressions :** To retrieve forest structure variables from VHR imagery, we relied on linear regressions between each of the 7 typical forest structure variables to be estimated and various textural features (first order and Haralick's features). The Pearson correlation coefficient  $R^2$  is used to analyze the correlation between each forest variable and texture features. Each forest variable is modeled as a function of a single texture feature. The best model for each variable is obtained automatically by selecting the one with the highest  $R^2$ .

**Towards a multi-scale approach :** The novelty in this approach is to combine features at different spatial and spectral resolutions using Pan and MS bands, and requiring different parametrizations, in an automated way. To achieve this goal, multiple regressions are used. Considering a large number of features, that are likely correlated, the challenge is in minimizing the multicollinearity on subset solutions in order to generate stable models and avoid overfitting. This can be done using the Variance Inflation Factor (VIF), a good multicollinearity detector which is defined as follows for any variable  $j$  involved in multiple regression :

$$v_j = \frac{1}{1-R_j^2}$$
 where  $R_j^2$  is the correlation coefficient from a regression of predictor  $j$  against the remaining predictors. The higher  $v_j$  (VIF), the higher the collinearity between the predictor  $j$  and the remaining predictors is high. VIF equals to 1 if there is no correlation between features. A critical value of 4 is usually used (Castillo et al., 2010).

Different subset selection methods exist for multiple regressions. In this study, we compared the performances of three methods : the classic step-by-step forward method, a LASSO (Least Absolute Shrinkage and Selection Operator) method, and a forward stepwise approach based on LARS (Least Angular Regression) (Efron et al., 2004). The LARS algorithm is a novel forward stepwise method. Unlike the classic Forward method which adds the variable leading to the highest F-statistic, the LARS algorithm adds the variable that better explains the current residuals. The LASSO minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. LASSO and LARS stepwise have the same computational costs as ordinary least square regression. To compare the models performances, the PRESS (Predicted Residual Error Square Sum) (Allen, 1974) is used. This criterion is defined as :

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i(-i))^2$$
 where  $\hat{y}_i(-i)$  is the prediction for observation  $i$  when this observation is used as a supplementary data for the corresponding regression. This criteria is equivalent to a leave-one-out cross validation : the prediction error for an observation is obtained with a model constructed without this observation.

In addition, the Mallows's  $C_p$  statistic is used to decide the appropriate number of variables to keep in the model. This statistic's expression is given by:

$$C_p = \left( \frac{SSE_p}{SSE^2} \right) - (n - 2p)$$
 where  $SSE^2$  is the mean square error of the full model, and  $SSE_p$  is the residuals square sum of the model containing the  $p$  variables of interest. The stop condition is when  $C_p = p$ , where  $p$  is the number of variables (Efron et al., 2004).

The subset selection method that minimizes the VIF will be preferred. We aim to fully automatically find a good subset solution without any manual intervention.

## 4 RESULTS AND DISCUSSION

All image features extractions are based on the use of the OrfeoToolbox library (OTB<sup>1</sup>) (Inglada and Christophe, 2009), statistical analysis is based on Rproject. The whole is managed using Python programming language.

### 4.1 Single solutions

The tested dataset is made up of 2894 and 4760 feature descriptors for Pan and Ms respectively. For each forest variable, numerous significant linear relationships with image features are found

<sup>1</sup>OTB: OrfeoToolbox <http://orfeo-toolbox.org/otb/>

(considering a p-value  $< 0.01$  for  $R^2$ ). No anomaly was observed on the normal hypothesis for those features. Features with the highest absolute correlation coefficient are presented in Table 3 for Pan and MS bands separately. All model parameters estimations are significant considering a 1% level.

Variable	Band	Feature	r	d	0	moment	R <sup>2</sup>	pvalue	RMSE
Cd	Pan	inertia	10	6	45	M	0.943	1.515e-07	0.342
	PIR	inertia	3	2	90	M	0.940	1.924e-07	0.350
Nah	Pan	inertia	25	1	90	M	0.943	1.459e-07	75.282
	B	corr	6	3	0	V	0.970	5.350e-09	54.153
Sp	Pan	inertia	25	1	45	M	0.928	4.881e-07	0.388
	PIR	invdif	6	3	0	M	0.926	5.390e-07	0.392
Ht	Pan	inertia	25	1	45	M	0.897	2.901e-06	1.428
	PIR	invdif	6	3	0	M	0.923	6.636e-07	1.233
Dbh	Pan	inertia	25	6	45	M	0.905	1.911e-06	0.027
	PIR	invdif	12	3	90	M	0.930	4.257e-07	0.023
Ba	Pan	inertia	25	7	45	V	0.610	0.002	2.567
	NDVI	corr	3	1	0	M	0.641	0.001	2.463
Age	Pan	invdif	25	6	45	M	0.893	3.564e-06	4.106
	G	variance	12			M	0.848	2.095e-05	4.892

Table 3: Forest structure variables and their best image descriptors

The best descriptors were found for density and crown diameter (which have the biggest impact on image texture) which are the Blue band correlation and the Pan inertia respectively. All other forest variables are well described except for the basal area. This result was expected considering literature (Kayitakire et al., 2006). Performances are similar for Pan and MS on this reduced set of stands. In addition, second order texture features (GLCM) clearly appear to be the most descriptive of forest structure, only one first order texture (green band variance) has been selected for age estimation. Inertia and inverse difference are the most frequently selected GLCM texture features in this study. The near infra red is the most relevant band for MS. A window radius of 25 is the best for Pan, it varies between 3 and 12 for MS. The optimal displacement for Pan is either 6 or 1 pixel with a 45° orientation while for MS, the optimal displacement is of 3 pixels with a 0° or 90° orientation.

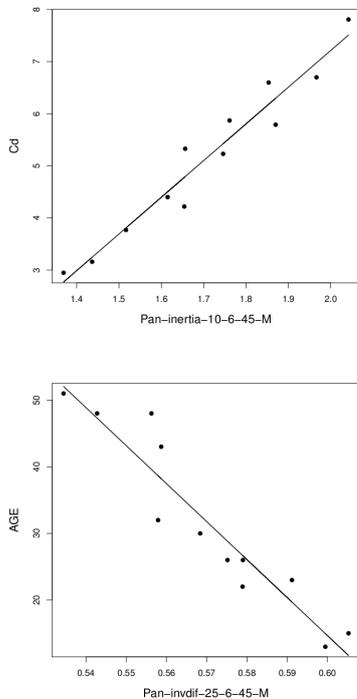


Figure 2: Relationship between forest structure variables and selected image features

Figure 2 shows some of the best achieved relationships between forest variables and image features. Linear presumptions are confirmed and no alarming behaviors are noticeable on residual plots.

The variability of feature parameters shows that a wide range of parameters has to be tested to optimize the prediction precision for a considered forest variable. Overall results are good compared to other works on VHR imagery. For instance, (Kayitakire et al., 2006) found the following correlation coefficients with other texture features (correlation and contrast) from the panchromatic Ikonos band (1m) and 30 observations of spruce stands: 0.81, 0.76, 0.82, 0.82 and 0.35 for age, Ht, Dbh, density and Ba. RMSE reached 2, 0.4 and 7 for Ht, Dbh and Ba.

One has to be cautious though not to draw fast conclusions from these results due to the limited data (12 observations) that has been used in the estimation process.

## 4.2 Multiple solutions

The determination of the best subset selection method was carried out by testing the three methods (classic Forward, Lasso and LARS stepwise) on all variables with both Pan and Ms datasets. Results are shown only for the estimation of age for which the highest number of observations is available (184 versus 12 for the other variables). The three methods behave similarly for all tested forest variables. The number of variables was fixed to three according to the Cp statistic. Figure 3 shows the Cp's curve (which indicates the gain when adding a variable at each step) for the LARS stepwise selections. The subset selection methods for re-

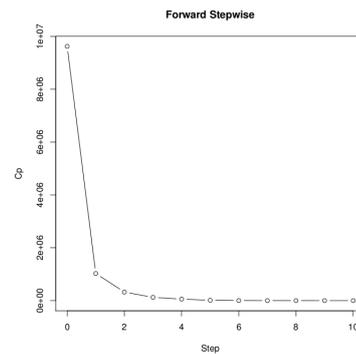


Figure 3: Cp's curve for LARS stepwise selection on panchromatic data retrieving the age variable are compared in Table 4 and Table 5 using Pan and MS data respectively. For each model, we show the LMG (initials of the authors who first proposed the method) statistic to estimate the contribution of each variable on the obtained model (Gromping, 2006). LMG averages the sum of sequential squares for all possible arrangements of the variables in the model. One can observe that the LARS stepwise selection generates a subset without multicollinearity (VIF close to 1). The forward selection leads to the best prediction performance considering the PRESS criteria, however VIF values are critical ( $>4$ ). The lasso selection proposes a more balanced model (considering the contribution of variables (LMG)) but the solution suffers from multicollinearity too.

Methods	Band	Feature	r	d	0	moment	LMG	VIF	PRESS
Forward	Pan	invdif	25	6	45	M	0.521	4.534	38.113
	Pan	clushade	25	5	90	M	0.393	2.779	
	Pan	energy	10	3	0	M	0.084	2.247	
Lasso	Pan	invdif	25	6	45	M	0.356	4.653	105.213
	Pan	inertia	20	8	0	V	0.320	6.247	
	Pan	clushade	20	6	90	M	0.322	3.725	
Stepwise	Pan	invdif	25	6	45	M	0.906	1.041	80.423
	Pan	inertia	25	3	45	M	0.078	1.069	
	Pan	corr	25	2	0	V	0.015	1.102	

Table 4: Comparison of subset selection methods in retrieving ages from panchromatic data ( $n = 12$ )

The LARS stepwise selection was then applied in a multi-scale approach, combining Pan and MS datasets, in order to find multiple solutions for each forest variable. Table 6 shows the best subset for all forest variables. The multiple regression improves the

Methods	Band	Feature	<i>r</i>	<i>d</i>	<i>0</i>	<i>moment</i>	LMG	VIF	PRESS
Forward	G	variance	12			M	0.434	7.738	116.77
	G	clushade	3	1	0	V	0.393	5.299	
	PIR	clushade	6	2	90	M	0.351	3.794	
Lasso	G	variance	12			M	0.336	5.365	361.283
	NDVI	clushade	3	2	45	V	0.332	5.612	
	R	corr	6	2	0	V	0.330	5.082	
Stepwise	G	variance	12			M	0.819	1.118	115.731
	PIR	corr	6	2	45	M	0.038	1.039	
	R	corr	9	4	0	V	0.142	1.081	

Table 5: Comparison of subset selection methods in retrieving ages from multispectral data ( $n = 12$ )

quality of predictions in all cases including the basal area compared to single regressions (Table 3). One can observe that most

	Band	Feature	<i>r</i>	<i>d</i>	<i>0</i>	<i>moment</i>	LMG	VIF	PRESS	RMSE
Cd	Pan	inertia	10	6	45	M	0.906	1.018	0.219	0.09
	G	corr	9	3	90	V	0.053	1.139		
	NDVI	harcorr	12	1	0	V	0.040	1.144		
Nah	B	corr	6	3	0	V	0.862	1.728	2640.916	10.875
	B	corr	6	3	45	M	0.086	1.968		
	Pan	corr	20	8	45	V	0.051	1.602		
Sp	Pan	inertia	25	1	45	M	0.928	1.247	0.254	0.09
	Pan	corr	5	3	90	V	0.033	1.110		
	Pan	corr	10	5	45	M	0.037	1.130		
Ht	PIR	invdif	6	3	0	M	0.871	1.124	5.887	0.402
	R	corr	9	4	0	V	0.114	1.061		
	PIR	corr	3	2	45	M	0.014	1.073		
Dbh	PIR	invdif	12	3	90	M	0.830	1.111	0.001	0.005
	B	corr	3	2	0	V	0.157	1.110		
	Pan	inertia	15	1	45	V	0.011	1.003		
Ba	NDVI	corr	3	1	0	M	0.786	1.134	10.859	0.487
	Pan	inertia	25	2	0	V	0.203	1.371		
	NDVI	clushade	12	3	0	M	0.009	1.234		
Age	Pan	invdif	25	6	45	M	0.841	1.046	84.538	1.287
	Pan	inertia	25	3	45	M	0.079	1.011		
	G	corr	12	3	90	V	0.079	1.056		

Table 6: Multiple solutions combining panchromatic and multispectral data ( $n = 12$ )

of the optimal subsets use jointly Pan and MS data which confirms the relevance of a multi-scale approach. Besides, models are composed of features with various parameterizations. The VIF exhibits again a very low multicollinearity (close to 1) between all the selected subsets of texture features, for every estimated forest variable. Hence, the subset selection method generated a relevant solution, ensuring both high accuracy and low multicollinearity, in all tested cases. Compared to similar works on multi-linear regressions (Wolter et al., 2009), our approach tackles the multicollinearity problem in a more convincing way, being both a non parametric approach (no threshold to set) and completely automatic on both the selection of the subset of texture features and the underlying optimal parameters (window size, displacement length and orientation). We chose to limit the range of Haralick parameters to experiment as a trade-off between accuracy and efficiency, thus leading to a sub-optimal solution for all estimated forest variables. A broader range of descriptors could be easily considered to improve accuracy but at higher computational costs.

### 4.3 Validation on stand age

All features were calculated following the same protocol for all the site stands as for the previous test frame. A square plot (120m) was defined around the center of each stand. The optimal single features for age prediction are presented in Table 7. Stand

Band	Feature	<i>r</i>	<i>d</i>	<i>0</i>	<i>moment</i>	$R^2$	pvalue	RMSE
Pan	invdif	10	8	45	V	0.706	< 2.2e-16	5.665
G	variance	3			M	0.663	< 2.2e-16	6.068

Table 7: Best image features for age prediction ( $n = 184$ )

age varies from 4 to 51 years. The introduction of observations with more variability leads to linear relationships of lower significance. Comparing these results with those obtained by 12 observations, we can notice that the same best features are found for both Pan and MS data (Table 3). The obtained window radius is lower, which could be explained by the presence of very young stands, corresponding to smaller texture structure. This shows that window size is very sensitive to tree size or tree spacing variations. The multi-scale approach combines features with

different parameters (especially window size and resolution, both strongly related to the intrinsic nature of the texture which could be more or less captured and retrieved depending on the set up of these important spatial parameters). It is a good way to capture a well adapted information. The LARS selection was then applied

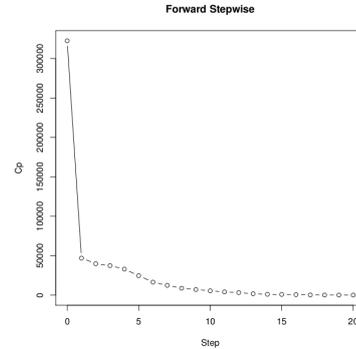


Figure 4: Cp's curve for age features selection

on the full dataset. The obtained variable subset is presented in Table 8 and the corresponding Cp's curve in Figure 4. The optimal feature subset is not subject to multicollinearity, confirming the efficiency of the LARS subset selection. The predicted RMSE (obtained by leave-one-out validation) is 5.25 years, corresponding to 10% of the age dynamic over the site.

Band	Feature	<i>r</i>	<i>d</i>	<i>0</i>	<i>moment</i>	LMG	VIF	PRESS	RMSE
Pan	invdif	10	8	45	V	0.510	2.551	3423.509	4.313
Pan	harcorr	5	1	135	M	0.123	1.357		
NDVI	harcorr	12	4	0	M	0.067	1.697		
Pan	inertia	25	2	45	V	0.011	1.150		
G	corr	3	4	135	M	0.080	1.353		
NDVI	corr	9	2	135	M	0.054	1.798		
Pan	corr	5	2	45	V	0.043	1.276		
PIR	corr	9	1	45	M	0.134	1.276		
R	corr	6	3	45	M	0.017	1.042		
NDVI	corr	6	4	90	V	0.088	1.075		

Table 8: Optimal image feature subset for age retrieval

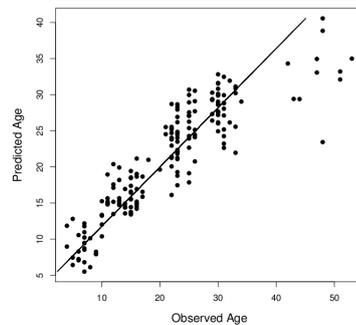


Figure 5: Observed and predicted age on the whole site

The predicted variable is relevant for stands younger than 35 years old (Figure 5).

In pine maritime forest of southwestern France, the oldest stands (over 40 years) are often heterogeneous and their canopy cover is often open mainly due to the silvicultural rules applied before the 70's. Moreover, above a given age (about 35 years), the trees growth slows down and the forest variables tend to be plateauing with age. This explains why the correlation between observed and predicted age decreases for the oldest stands.

In order to test the feasibility of a true stand age prediction, we split the data into two populations (which both have the same distribution) and computed the RMSE variation obtained when predicting the second age population with a sub-sample of the first population. Varying the size of the learning age population set from 20 observations to 80 (with 100 experiments) for the prediction of the whole validation dataset ( $n = 92$ ) (Figure 6), we

obtained an averaged RMSE ranging from 9 to 7.5 years. As we introduce a lot of randomness, the RMSE's standard deviation is important. This highlights the strong influence of the choice of observations and their number on the relevance of the test frame construction.

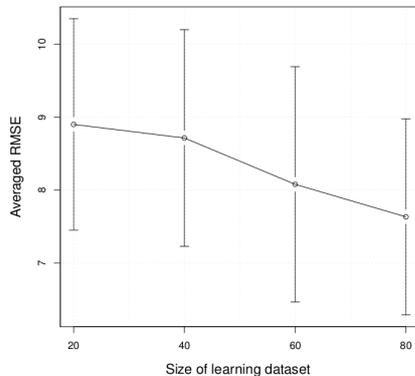


Figure 6: RMSE variations of predicted age as a function of the observation number (each point is the average of 100 experiments)

Finally, as crown diameter and spacing are strongly correlated with age we can expect for these forest variables a similar behavior regarding the statistical analysis performances presented in this section.

## 5 CONCLUSIONS AND FUTURE WORK

This study provided an automatic method to retrieve forest structure variables using spectral and Haralick's texture extracted from VHR optical satellite images. An automatic feature selection process whose originality lies on the use of image test frames of adequate forest samples was tested. This method allows a fast assessment of thousands of descriptors, exploring a wide range of parameter values. The best image features are selected via statistical modelling.

Seven typical forest structure variables were successfully modeled. Apart from the basal area, the regression coefficient,  $R^2$ , of the best single models ranges from 0.89 to 0.97. Then, a multi-scale approach, combining panchromatic and/or multi-spectral features with different parameterizations, is proposed. The multicollinearity problem is addressed carefully using the VIF criterion. Comparing three variable subset selection methods, a new stepwise method, derived from LARS, turned out the most convincing, significantly improving the quality of estimation for all the forest structure variables, including the basal area ( $R^2 > 0.98$ ). A validation on stand age retrieval over the whole site highlighted the potential of a multi-scale approach for retrieving forest structure variables from VHR satellite images. The whole protocol we have introduced can be easily applied to any other forest site data using site-specific test frames.

Our method will be applied on different sites, using other VHR sensors, expecting data from the new Pleiades satellite which has been successfully launched in December 2011.

## ACKNOWLEDGEMENTS

This research was supported by the regional council of Bordeaux (Région Aquitaine) and CNES<sup>2</sup> through the ORFEO program. The authors acknowledge INRA<sup>3</sup> for the VHR imagery and corresponding field data.

<sup>2</sup>Centre National d'Etudes Spatiales

<sup>3</sup>Institut National de Recherche Agronomique

## REFERENCES

- Allen, D. M., 1974. The relationship between variable selection and prediction. *Technometrics* 16:125, pp. 125–127.
- Barbier, N., Couteron, P., Proisy, C., Malhi, Y. and Gastellu-Etchegorry, J. P., 2010. The variation of apparent crown size and canopy heterogeneity across lowland amazonian forests. *Global Ecology and Biogeography* 19, pp. 72–84.
- Castillo, M. A., Ricker, M. and Jong, B. H. J. D., 2010. Estimation of tropical forest structure from spot-5 satellite images. *International Journal of Remote Sensing* 31-10, pp. 2767–2782.
- Coburn, C. A. and Roberts, A. C. B., 2004. A multiscale texture analysis procedure for improved forest stand classification. *International Journal of Remote Sensing* 25, pp. 4287–4308.
- Efron, B., Hastie, T. and Tibshirani, I. J. R., 2004. Least angle regression. *The Annals of Statistics* 32-2, pp. 407–499.
- Feng, Y., Li, Z. and Tokola, T., 2010. Estimation of stand mean crown diameter from high-spatial-resolution imagery based on a geostatistical method. *International Journal of Remote Sensing* 31-2, pp. 363–378.
- Franklin, S. E., Maudie, A. J. and Lavigne, M. B., 2001. Using spatial co-occurrence texture to increase forest structure and species composition classification accuracy. *Photogrammetric Engineering & Remote Sensing* 67-7, pp. 849–855.
- Gromping, U., 2006. Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software* 17, pp. 1–27.
- Guyon, D. and Riom, J., 1996. Estimation de caractéristiques forestières à partir d'images à haute résolution spatiale (spot5). *BUL. S.F.P.T.* 141, pp. 46–50.
- Haralick, R. M., Shanmugan, K. and Dinstein, I., 1973. Texture features for image classification. *IEEE Transactions Systems, Man and Cybernetics* pp. 610 – 621.
- Huang, C., Song, K., Kim, S., Townshend, J. R. G., Davis, P., Masek, J. G. and Goward, S. N., 2008. Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment* 112(3), pp. 970–985.
- Hyypä, J., Hyypä, H., Inkinen, M., Engdahl, M., Linko, S. and Zhu, Y. H., 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stands attributes. *Forest Ecology and Management* 128, 1-2, pp. 109–120.
- Inglada, J. and Christophe, E., 2009. The orfeo toolbox remote sensing image processing software. In: *IEEE International Geoscience and Remote Sensing Symposium*.
- Kayitakire, F., Hamel, C. and Defourny, P., 2006. Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery. *Remote Sensing of Environment* 102, pp. 390–401.
- M.Tuceryan, 1998. *The Handbook of Pattern Recognition and Computer Vision* (2nd Edition)(207-248). C. H. Chen, L. F. Pau, P. S. P. Wang.
- Murray, H., Lucieer, A. and R.Williams, 2010. Texture-based classification of sub-antarctic vegetation communities on heard island. *International Journal of Applied Earth Observation and Geoinformation* 12, pp. 138–149.
- Ozdemir, I. and Karnieli, A., 2011. Predicting forest structural parameters using the image texture derived from worldview-2 multispectral imagery in a dryland forest, israel. *International Journal of Applied Earth Observation and Geoinformation* 13, pp. 701–710.
- Pesari, M., Gerhardinger, A. and Kayitakire, F., 2008. A robust built-up area precense index by anisotropic rotation-invariant textural measure. *IEEE Journal of selected topics in applied earth observations and remote sensing*.
- Proisy, C., Couteron, P. and Fromard, F., 2007. Predicting and mapping mangrove biomass from canopy grain analysis using fourier-based textural ordination of ikonos images. *Remote Sensing of Environment* 109, pp. 379–392.
- Song, C., Dickinson, M. B., Su, L., Zhang, S. and Yaussey, D., 2010. Estimating average tree crown size using spatial information from ikonos and quickbird images: Across-sensor and across-site comparisons. *Remote Sensing of Environment* 114, pp. 1099–1107.
- Wolter, T., Townsend, P. A. and Sturtevant, B. R., 2009. Estimation of forest structural parameters using 5 and 10 meter spot-5 satellite data. *Remote Sensing of Environment* 113, pp. 2019–2036.
- Wunderle, A. L., Franklin, S. E. and Guo, X. G., 2007. Regenerating boreal forest structure estimation using spot-5 pansharpened imagery. *International Journal of Remote Sensing* 28-19, pp. 4351–4364.



## Annexe D

**Modelling-based feature selection for classification of forest structure using very high resolution multispectral imagery.**

# Modelling-based feature selection for classification of forest structure using very high resolution multispectral imagery

Benoît Beguet<sup>1,2</sup>, Samia Boukir<sup>1</sup>G&E Laboratory (EA 4592)IPB / University of Bordeaux  
Bordeaux, France

name.surname@ipb.fr

Dominique Guyon

<sup>2</sup>UR 1263 EPHYSE

INRA

Bordeaux, France

dominique.guyon@bordeaux.inra.fr

Nesrine Chehata<sup>3,1</sup><sup>3</sup>UMR LISAHINRA-IRD-SupAgro  
Tunis, Tunisia

nesrine.chehata@ipb.fr

**Abstract**—This paper presents a new feature selection method which aims to effectively and efficiently map remote sensing data. An automated texture-based modelling procedure of forest structure variables is at the core of our approach. We show that texture features that are highly correlated to genuine physical parameters of forest structure have potential for building reliable classifiers. We demonstrate the effectiveness of our modelling-based texture feature selection method in performing mapping of very high resolution forest images. Our method outperforms Random Forest variable importance in terms of classification accuracy and computational complexity.

**Index Terms**—Feature selection, Modelling, Classification, Texture, Forest

## I. INTRODUCTION

Classification of remotely sensed data has long attracted the attention of the remote-sensing community. Scientists and practitioners have made great efforts in developing advanced classification approaches and in improving classification accuracy and speed [17]. In remote sensing, numerous features can be derived from original spectral bands. This significantly increases the computational complexity of remote sensing classification tasks which handle massive information, especially if texture features are involved.

Feature selection is a long-studied topic. It can be used as a preprocessing step for classification to provide the most relevant attributes as input features. It not only reduces classification complexity but can also increase classification accuracy especially for minority classes [12], [8]. In remote sensing, feature selection prior to classification is usually used in the context of high-dimensional data. It could be hyper-spectral image classification [16], [22] or like in our case, texture analysis [7], [21]. Mainly three different types of feature selection methods exist in literature [12]: filters, wrappers and embedded methods. Filters are independent of the classifier, exploring only the intrinsic discriminative power of features. These methods are fast but suboptimal. Wrappers use the classification algorithm as an evaluation mean. They are powerful but induce very high computational costs. Embedded methods include the feature selection in the learning process. This

allows a better use of the data because it does not need a learning step for each candidate feature subset.

An efficient *embedded* feature selection method involves the variable importance produced by Random Forests (RF), a powerful decision tree-based ensemble method [3], [10], [20]. Considering the bagging family method as an example, the importance of a feature  $f$  can be estimated by randomly permuting all the values of this feature in the OOB (Out Of Bag) samples for each tree. OOB refers to the instances that have not been picked up in bagging for bootstrapped training samples. For other ensemble methods, OOB can be replaced by a set aside data set. The RF variable importance follows the idea that a random permutation of a feature mimics the absence of that feature from the model. The importance of a feature  $f$  is measured as the difference between the prediction accuracy before and after permuting feature  $f$ , averaged over all trees [3]. We have successfully used this *permutation importance* measure for urban mapping [11]. However, feature selection via supervised classification is not only costly but also dependent of the underlying classifier.

We investigate in this paper the potential of feature selection derived from a modelling process for classifier design in the difficult context of forest remote sensing. The main novelty of our work is that the feature selection criterion is based on the best correlation between features and measured physical parameters. Hence, the underlying selected features have a clear physical meaning that can contribute to a better understanding of the classification task, thus leading to a more robust classification. This approach has an appealing property: *it is independent of classifiers*. Furthermore, it is much faster than state-of-the-art feature selection techniques for classification which generally involve a machine learning algorithm.

This paper is organized as follows. The following section describes our new modelling-based texture feature selection method. The validation of our approach on forest cover classification from Very High Resolution (VHR) imagery is presented in section III. Discussions and concluding remarks are given in the last section.

## II. A NEW MODELLING-BASED FEATURE SELECTION METHOD

An innovative feature selection algorithm is proposed to effectively and efficiently map forest structure. The main novelty of our approach is in the use of physical parameters in the feature selection process. This strategy allows to significantly reduce the computational complexity of this step prior to classification. And, more importantly, it provides suitable features as inputs to any classifier whether it is supervised or not. Hence, it is a viable alternative to more traditional feature selection methods [12], [8] and has potential to better address the complexity of forest mapping.

Our feature selection is a non parametric approach (no threshold to set) based on statistical modelling, exploiting textural information from VHR multispectral imagery. The best texture features are selected using a multiple linear regression method based on a Least Angle Regression scheme (LARS) [15]. It fully automatically finds a relevant set of texture features that best describe forest structure without any manual intervention. Indeed, it is completely automated on both the selection of the subset of texture features and optimization of the underlying parameters (window size,...). This is achieved thanks to a fast supervised scheme involving test frames (see fig. 2).

Figure 1 shows a flow chart of our feature selection scheme for forest structure classification. It consists of three main steps described below:

- Automated image texture analysis via test frames
- Automated forest structure variable retrieval
- Forest variable texture feature combination

### A. Automated image texture analysis via test frames

Texture analysis of VHR forest images mainly relies on GLCM (Grey Level Co-occurrence Matrix) features [14], [5], [6], [24], [19]. GLCMs are second order statistics constructed by 2D histograms [13]. The main drawback of this well-known texture representation is the underlying parameters (window size, displacement length, orientation, and quantification level) which are extremely difficult to set, especially in forestry due to the spatial complexity of the forest structure. To tackle this major issue, an automated process that explores a wide range of texture parameter values is carried out here. The originality of this procedure, inspired by camera calibration protocols, lies on the use of a test frame of adequate forest image samples (see fig. 2). Hence, just a few small image samples are required to build up the test frame but allow a fast assessment of hundreds of descriptors, given the large number of tested combinations of texture parameter settings. Compared to more traditional supervised methods, which are very costly in terms of field surveys and time-consuming, our method requires a slight supervision. Furthermore, in the context of texture analysis of VHR images, which induces high computational costs, the use of test frames appears as a more practical alternative than a traditional machine learning step.

We consider first order texture features (mean and variance) and second order GLCM texture features. Haralick et

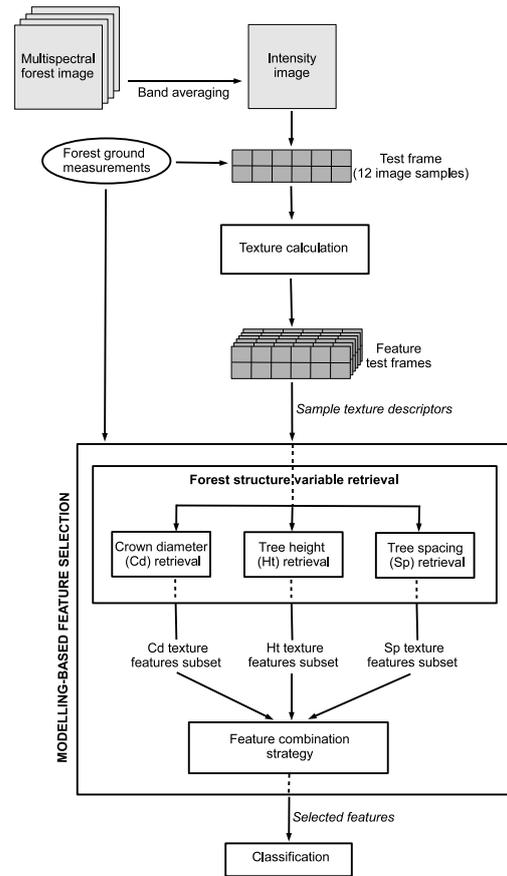


Fig. 1. Feature selection scheme for forest mapping

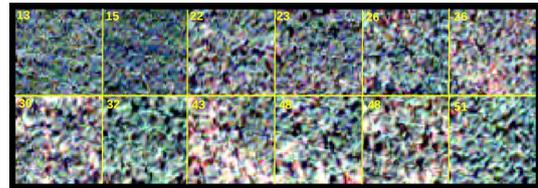


Fig. 2. Multispectral test frame of 12 forest samples, of size  $50 \times 50$  pixels, sorted by increasing tree age

al. defined 14 texture features derived from GLCM [13]. We use the eight most used Haralick's features in remote sensing forestry [14], [5]: energy, entropy, inertia, homogeneity (inverse differences), correlation, Haralick's correlation, cluster shade and cluster prominence. GLCM parameters are: radius  $r$  of the moving window (window size =  $2r + 1$ ), displacement  $d$ , orientation  $o$  and quantification number  $nbbin$ . Texture analysis is carried out on an *intensity* image resulting from a uniform averaging of the available multispectral bands.

### B. Automated forest structure variable retrieval

Many studies have focused on estimating forest variables from remote sensing data, exploiting spectral or textural image

features [5], [23], [19], [7]. We propose an automated modelling process of forest structure using test frames of adequate forest image samples, previously introduced, where the structure variables were measured at ground (fig. 2). In order to identify correlation between candidate image texture features (first order and second order GLCM features) and forest structure variables, we use a statistical modelling approach. More exactly, a LARS (Least Angle Regression) stepwise multiple linear regression technique [15] is carried out to estimate three typical forest structure variables: Crown diameter (Cd), tree Height (Ht) and tree Spacing (Sp). The inputs of this process are provided by the preceding image texture analysis step. The latter produces a large number  $n = 8n_p + 2n_w$  ( $n_p$  = number of GLCM parameter settings and  $n_w$  = number of window size settings) of texture test frames, considering 8 second order GLCMs and 2 first order texture features. For each texture calculation of the twelve forest image samples, a single representative value is chosen as the median value of the  $50 \times 50$  pixel sample texture result. We chose median values instead of mean values, which are traditionally used for this matter [2], [14], [5], in order to alleviate smoothing effects. Hence,  $n$  different texture descriptors, associated to each forest image sample, are used as inputs in the modelling process.

In multiple linear regression, considering a large number of features that are likely correlated, the challenge is in minimizing multicollinearity on subset solutions in order to generate stable models and avoid over-fitting. This can be done with the Variance Inflation Factor (VIF), an efficient multicollinearity detector [5] which is defined as follows, for any variable  $j$  involved in multiple regression:

$$v_j = \frac{1}{1 - R_j^2} \quad (1)$$

where  $R_j^2$  is the regression coefficient from a regression of predictor  $j$  against the remaining predictors. The higher  $v_j$  (VIF), the higher the collinearity between predictor  $j$  and the remaining predictors. VIF equals to 1 if there is no correlation between features [5].

Different subset selection methods exist for multiple regression. In previous work [2], we compared the performances of three methods: the classic step-by-step forward method, a LASSO (Least Absolute Shrinkage and Selection Operator) method, and a LARS stepwise approach [15]. Unlike the classic forward method which iteratively adds the variable leading to the highest F-statistic, the LARS algorithm adds the variable that better explains the current residuals. The LASSO algorithm minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. To compare model performances, the PRESS (Predicted Residual Error Square Sum) criterion [1] was used. It is defined as :

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

where  $\hat{y}_i(-i)$  is the prediction for observation  $i$  when this observation is used as a supplementary data for the corresponding

regression. This criterion is equivalent to a LOO (Leave-One-Out) cross-validation. In addition, the Mallows'  $C_p$  statistic [15] was used to decide the appropriate number of variables to keep in the model. The subset selection method that minimizes VIF was preferred, hence favouring as uncorrelated variables as possible in the selected subset. LARS stepwise procedure turned out the most effective subset selection method [2] and hence, it is the one involved in our modelling-based texture feature selection method. Furthermore, it has as low complexity as ordinary least square regression.

### C. Forest variable texture feature combination

The preceding modelling stage provides an ordered subset of texture features of decreasing importance for each forest structure variable (Crown diameter Cd, tree Height Ht, tree Spacing Sp). A strategy to combine these 3 different sets of features is needed to further reduce classification complexity and eventually increase classification accuracy. In a way, this final step is also a feature selection issue that is of particular significance in the difficult context of forest structure mapping. Indeed, due to allometric constraints, a strong correlation is observed between Cd, Ht and Sp [2]. Therefore, one has to be cautious when attempting to combine the feature subsets associated to each forest structure variable. In this work, we use two simple strategies to derive a global forest descriptor from the 3 forest variable feature subsets. More sophisticated feature combination schemes will be investigated in future work to better capture the forest structure complexity. The two feature combination strategies consist respectively in:

- Combining the best texture feature of each of the forest structure variables (Cd, Ht and Sp), leading to a 3-feature forest descriptor involving all three forest structure variables.
- Using the 3 best texture features of exclusively one of the three forest structure variables (Cd or Ht or Sp), leading to a forest descriptor of size 3 too but involving solely one forest structure variable.

The last strategy is worthwhile to investigate because of the strong correlation existing between forest structure variables. Furthermore, it is of particular interest from a thematic point of view. Indeed, some forest structure variables are more difficult to measure than others and have higher measure uncertainty. For example, Crown diameter (Cd) is more troublesome than tree Height (Ht) and hence, leads to a lower classification performance as shown later in our experiments (section III-C).

Let us notice that the preceding forest structure variable modelling step provided, for all considered forest variables, an optimal feature size  $s = 3$ , using our forest data, based on Mallows'  $C_p$  statistic [15]. This suggests that about 3 features are enough to estimate the forest structure variables. However, this does not necessarily mean that this feature size is also sufficient for the classification task. This important issue will be further investigated in the future.

Finally, for performance comparative analysis, we use the RF (Random Forest) permutation importance method. Both feature selection methods are validated on two well-established

supervised classifiers: Bagging Trees (BT) [4] (to be consistent with decision tree-based feature selection) and Support Vector Machine (SVM), another powerful classifier, widely used in remote sensing [18].

### III. EXPERIMENTAL RESULTS

#### A. Remote sensing and field data

Our method has been applied to a multispectral Quickbird image (acquired in 2003) of a maritime pine forest in South-West France (fig. 3 exhibits a sample of this image). This 2.4 m spatial resolution image combines 4 spectral bands: Red (R), Green (G), Blue (B) and Near-Infra-Red (NIR).

Measurements of forest structure variables, tree Height (Ht), crown Diameter (Cd) and tree Spacing (Sp), are available on 12 stands of varying tree age (see fig. 2 for an image sample of each stand).

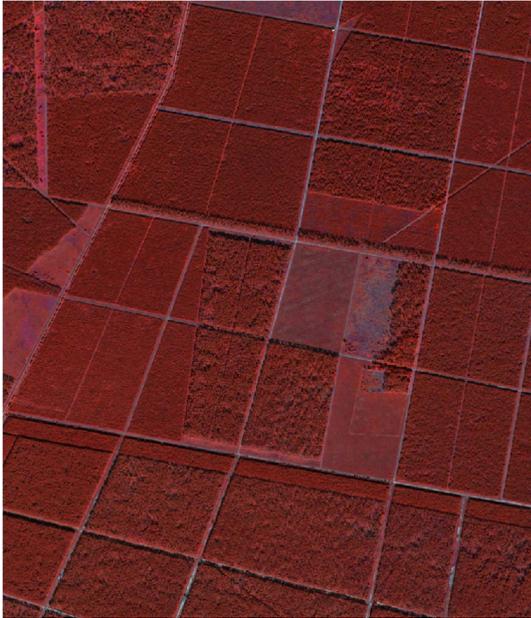


Fig. 3. *Quickbird* multispectral image sample (ROI 2), of size  $1136 \times 972$  pixels, of *Nezer* forest (R, G, NIR)

In image texture analysis, we chose to limit the range of Haralick's parameters to achieve a trade-off between accuracy and efficiency (table I). Hence, a total of  $n = 8n_p + 2n_w = 520$  ( $n_p$  = number of GLCM parameter settings and  $n_w$  = number of window size settings) texture features, of varying texture parameter settings, is experimented. As for the quantification level, since only forest pixels (in image samples) are considered, setting  $n_{bbin}=8$ , as traditionally used in literature, appears to be sufficient.

$r$ (pixel)	$d$ (pixel)	$o$ (degree)
3-12 step: 3	1-4 step:1	0-135 step: 45

TABLE I  
GLCM PARAMETER RANGE VALUES

1) *Test frames*: The forest test frame (figure 2) is constructed with 12 small image samples associated with field measurements on Cd, Ht and Sp, which vary respectively from 2.9m to 7.8m, 9.2m to 24.6m, and 3m to 7.8m. Each of the 12 image samples, of size  $50 \times 50$  pixels, was selected at the centre of corresponding field measurement area.

2) *Datasets*: Five forest structure classes have been designed to describe the whole forest structure diversity along the site ( $10 \times 4$ km area). Class 1 denotes areas with no visible trees in the image (Cd less than 0.5m). Classes 2-5 correspond to different forest structure growth stages, from very regular in-line stands with high tree density and low canopies to very heterogeneous stands with low tree density and high canopies. Two distinct ROIs (Region Of Interest) of size respectively  $1109 \times 1050$  pixels and  $1136 \times 972$  pixels (fig. 3), extracted from the Quickbird forest image, are tested. Only *forest* pixels were included in datasets, the used mask and reference stands were provided as a GIS layer. Of course, these ROIs are independent of the test frame (depicted on fig. 2) involved in the modelling-based feature selection process. Each of the two datasets has been divided into two equal parts, training set and test set, for supervised classification.

#### B. Feature selection via statistical modelling

For each forest structure variable, we show, on table II, the selected texture features and the underlying optimal texture parameter settings (window radius  $r$ , displacement  $d$  and orientation  $o$ ) as well as three statistical criteria: VIF (Variance Inflation Factor), PRESS (Predicted Residual Error Square Sum) and RMSE (Root of Mean Square Error). For every forest variable, the statistical modelling method generated a relevant solution, ensuring both high prediction performance and low multicollinearity (VIF close to 1).

Forest variable	Texture feature	$r$ (pix.)	$d$ (pix.)	$o$ (deg.)	VIF	PRESS	RMSE
Cd	inertia	3	2	90	1.04	0.70	0.16
	cl. shade	3	2	135	1.08		
	corr.	3	2	135	1.12		
Ht	inv. dif.	6	3	0	1.36	10.76	0.63
	corr.	9	2	45	1.06		
	cl. shade	12	3	0	1.31		
Sp	inv. dif.	9	3	0	1.03	1.11	0.21
	cl. shade	3	2	135	1.02		
	corr.	12	2	45	1.02		

TABLE II  
MODELLING-BASED TEXTURE FEATURE SELECTION

#### C. Feature selection performance comparative analysis

Our modelling-based feature selection method is compared to RF (Random Forest) permutation importance based on a bagging ensemble of 100 decision trees. Both feature selection methods use the same test frame. Our automated modelling procedure providing three best features for each forest variable, according to  $C_p$  statistic [15], we set the number of selected texture features to consider in classification, for both feature selection methods, to three. Hence, a descriptor of 7 features is considered for pixel classification: four spectral bands (R,G,B,NIR) and three texture features provided by the

feature selection step. The classification is carried out by a tree-based ensemble method, namely Bagging Trees (BT) with 100 trees, and SVM, based on a linear kernel.

1) *Overall classification performance*: The classification performances obtained using the 3 best texture features provided by RF permutation importance are compared to the performances achieved with the set of 3 best texture features provided by our modelling-based feature selection method, for respectively each forest variable (Cd, Ht, Sp). A descriptor combining the best feature for each forest variable is also considered (see section II-C). Table III shows the overall classification accuracies achieved for all sets of features and all classifiers.

ROI 1	RF imp.	Mod. Cd	Mod. Ht	Mod. Sp	Mod. Cd,Ht,Sp
BT	74.712	70.52	75.32	<b>76.54</b>	75.51
SVM	75.52	70.44	75.71	<b>76.47</b>	76.21
ROI 2	RF imp.	Mod. Cd	Mod. Ht	Mod. Sp	Mod. Cd,Ht,Sp
BT	82.43	77.43	81.95	83.33	<b>83.47</b>
SVM	<b>83.53</b>	78.03	82.17	83.10	83.23

TABLE III  
OVERALL CLASSIFICATION PERFORMANCES (%) USING DIFFERENT FEATURE SELECTION STRATEGIES ON TWO SAMPLES OF *Quickbird* MULTISPECTRAL FOREST IMAGE

These results demonstrate the relevance of our feature selection method for classification which achieve similar or even better classification accuracy than a major feature selection method: RF variable importance. The improvement in classification accuracy is up to 2% (BT on 3 best modelling-based texture features for Sp *versus* BT on 3 best features by RF permutation importance, for ROI 1).

Figure 4 presents both reference map and classification results for one of the considered samples of the *Quickbird* forest image (ROI 2), depicted on fig. 3. The displayed classification results are provided by Bagging Trees on 3 best modelling-based texture features for Sp (tree Spacing).

More field measurements will yield more desirable results in statistical modelling and hence in feature selection and classification outcome. Furthermore, more relevant combinations of best texture features associated to each forest variable could improve the classification results.

2) *Classification performance per class*: Table IV presents the maximum classification error rate per class obtained on test set by BT and SVM classifiers using the 3 best texture features provided by both our modelling-based and RF feature selection methods. This table clearly shows that our method largely outperforms the RF permutation importance method in terms of handling minority and complex classes. Indeed, they are obviously subject to the highest classification error rates. Except for the case of BT classification of ROI 1, for which the permutation importance slightly outperforms our method (maximum classification error of 36.82% *versus* 37.71%), the three other cases are all significantly in favour of our method. The achieved improvement is up to **32%** (BT on best modelling-based texture features for Cd, Sp and Ht *versus* BT on 3 best features by RF permutation importance, for ROI 2). These results demonstrate that the involvement of physical

properties of forest structure in image feature selection is an appealing solution for building efficient classifiers.

ROI 1	RF imp.	Mod. Cd	Mod. Ht	Mod. Sp	Mod. Cd,Ht,Sp
BT	<b>36.82</b>	61.89	39.84	37.71	39.52
SVM	62.65	73.79	39.47	<b>38.43</b>	45.03
ROI 2	RF imp.	Mod. Cd	Mod. Ht	Mod. Sp	Mod. Cd,Ht,Sp
BT	84.27	72.68	58.28	52.70	<b>52.06</b>
SVM	87.04	78.37	66.26	60.17	<b>60.13</b>

TABLE IV  
MAXIMUM CLASSIFICATION ERROR RATE PER CLASS (%) USING DIFFERENT FEATURE SELECTION STRATEGIES, FOR TWO SAMPLES OF *Quickbird* MULTISPECTRAL IMAGE

3) *Complexity*: Our modelling-based feature selection technique has a significantly lower time complexity than RF permutation importance. For a three feature subset, selected over a total of 520 features, the elapsed CPU time for RF feature selection is 466.77s while our feature selection is achieved in only 1.02s, a decrease in computational costs of almost a factor **500**. Hence, our approach achieves a good classification accuracy given the complexity of textured forest images (see fig. 2) with a simple and very fast feature selection scheme.

#### IV. CONCLUSION

We have shown that texture features that are highly correlated to genuine physical parameters of forest structure are an appealing choice for image classification. Our feature selection method, deriving from a statistical modelling procedure, is independent of classifiers and better addresses the computational load of remote sensing data. The estimation of forest structure variables from image texture is an automated process whose originality lies on the use of test frames of adequate forest image samples. This method allows a fast assessment of hundreds of descriptors, exploring a wide range of texture parameter settings.

Our method achieves similar or even better classification performance than Random Forest permutation importance while being a much faster alternative (time complexity divided by up to a factor **500**). Furthermore, it better handles minority and difficult classes, with an improvement in classification performance per class of up to **32%**. Hence, our technique is a viable alternative to traditional methods. It will be validated on another forest site with different VHR satellite images and measurements on more forest stands (representative of all forest structure classes).

#### ACKNOWLEDGMENT

This research was supported by the regional council of Bordeaux (Région Aquitaine) and CNES<sup>1</sup> through the ORFEO program. The authors acknowledge INRA<sup>2</sup> for the VHR imagery and corresponding field data. They also wish to thank Christian Germain from IMS Lab (Bordeaux), Jordi Inglada from CESBIO (Toulouse) and Jean-Pierre Wigneron from INRA (Bordeaux) for discussions.

<sup>1</sup>Centre National d'Etudes Spatiales

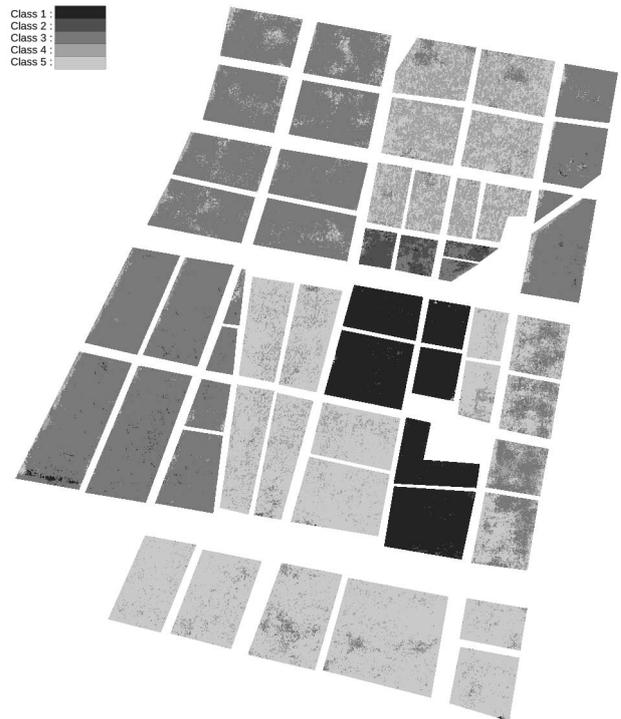
<sup>2</sup>Institut National de Recherche Agronomique

## REFERENCES

- [1] D. M. Allen, *The relationship between variable selection and prediction*. *Technometrics*, vol. 16:125, pp. 125–127, 1974.
- [2] B. Beguet, N. Chehata, S. Boukir, and D. Guyon, *Retrieving forest structure variables from very high resolution satellite images using an automatic method*. *ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences*, vol. I-7, pp. 1–6, 2012.
- [3] L. Breiman, *Random forests*. *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [4] L. Breiman, *Bagging predictors*. *Machine Learning*, vol. 24, pp. 123–140, 1996.
- [5] M. A. Castillo, M. Ricker, and B. H. J. De Jong, *Estimation of tropical forest structure from spot-5 satellite images*. *International Journal of Remote Sensing*, vol. 31-10, pp. 2767–2782, 2010.
- [6] C. A. Coburn and A. C. B. Roberts, *A multiscale texture analysis procedure for improved forest stand classification*. *International Journal of Remote Sensing*, vol. 25, pp. 4287–4308, 2004.
- [7] F.M.B. Van Coillie and Robert R. De Wulf L.P.C. Verbeke, *Feature selection by genetic algorithms in object-based classification of ikonos imagery for forest mapping in flanders, belgium*. *Remote Sensing of Environment*, vol. 110, pp. 476–487, 2007.
- [8] M. Dash and H. Liu, *Feature selection for classification*. *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.
- [9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. Wiley-Interscience, 2nd edition, October 2001.
- [10] R. Genuer, J.M. Poggi, and C. Tuleau-Malot, *Variable selection using random forests*. *Pattern Recognition Letters*, vol. 31, pp. 2225–2236, 2010.
- [11] L. Guo, N. Chehata, C. Mallet, and S. Boukir, *Relevance or airborne lidar and image data for urban scene classification using random forests*. *ISPRS Journal of Photogrammetry and remote Sensing*, vol. 66-1, pp. 56–66, 2011.
- [12] I. Guyon and A. Elissee, *An introduction to variable and feature selection*. *Journal of Machine Learning Research*, vol. 3(7-8), pp. 1157–1182, 2003.
- [13] R. M. Haralick, K. Shanmugan, and I. Dinstein, *Texture features for image classification*. *IEEE Transactions Systems, Man and Cybernetics*, pp. 610 – 621, 1973.
- [14] F. Kayitakire, C. Hamel, and P. Defourny, *Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery*. *Remote Sensing of Environment*, vol. 102, pp. 390–401, 2006.
- [15] B. Efron, T. Hastie, and I. J. R. Tibshirani, *Least angle regression*. *The Annals of Statistics*, vol. 32-2, pp. 407–499, 2004.
- [16] S. Li, H. Wu, D. Wan, and J. Zhu, *An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine*. *Knowledge-Based Systems*, vol. 24, pp. 4048, 2011.
- [17] D. Lu and Q. Weng, *A survey of image classification methods and techniques for improving classification performance*. *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.
- [18] G. Mountrakis, J. Im, and C. Ogole, *Support vector machines in remote sensing : A review*. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, pp. 247–259, 2011.
- [19] I. Ozdemir and A. Karnieli, *Predicting forest structural parameters using the image texture derived from worldview-2 multispectral imagery in a dryland forest, israel*. *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, pp. 701–710, 2011.
- [20] M. Pal, *Random forest classifier for remote sensing classification*. *International Journal of Remote Sensing*, vol. 26(1), pp. 217–222, 2005.
- [21] D. Puig, M.A. Garcia, and J. Melendez, *Application-independent feature selection for texture classification*. *Pattern Recognition*, vol. 43, pp. 3282–3297, 2010.
- [22] S. B. Serpico and L. Bruzzone, *A new search algorithm for feature selection in hyperspectral remote sensing images*. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39(7), pp. 1360–1367, 2001.
- [23] S. Tuominen and A. Pekkarinen, *Performance of different spectral and textural aerial photograph features in multi-source forest inventory*. *Remote Sensing of Environment*, vol. 94, pp. 256–268, 2005.
- [24] T. Wolter, P. A. Townsend, and B. R. Sturtevant, *Estimation of forest structural parameters using 5 and 10 meter spot-5 satellite data*. *Remote Sensing of Environment*, vol. 113, pp. 2019–2036, 2009.



(a) Reference map (5 forest structure classes)



(b) Tree-based ensemble classification using 3 best texture features for Sp (tree Spacing)

Fig. 4. Classification of *Quickbird* multispectral image sample (ROI 2) of *Nezer* forest using modelling-based feature selection



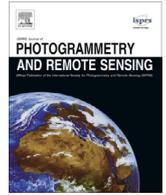
## Annexe E

Automated retrieval of forest structure variables based on multi-scale texture analysis of VHR satellite imagery.



Contents lists available at ScienceDirect

## ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: [www.elsevier.com/locate/isprsjprs](http://www.elsevier.com/locate/isprsjprs)

# Automated retrieval of forest structure variables based on multi-scale texture analysis of VHR satellite imagery



Benoit Beguet<sup>a,b,\*</sup>, Dominique Guyon<sup>b</sup>, Samia Boukir<sup>a</sup>, Nesrine Chehata<sup>a</sup>

<sup>a</sup> EA 4592 G&E, ENSEGD, University of Bordeaux, 1 Allée F. Daguin, 33607 Pessac Cedex, France

<sup>b</sup> INRA, UMR1391 ISPA, 33140 Villenave d'Ornon, France

## ARTICLE INFO

### Article history:

Received 5 February 2014

Received in revised form 28 April 2014

Accepted 8 July 2014

### Keywords:

Forestry  
Multiple regression  
Feature selection  
Texture  
Multi-scale  
Multi-resolution  
Pléiades  
Quickbird

## ABSTRACT

The main goal of this study is to design a method to describe the structure of forest stands from Very High Resolution satellite imagery, relying on some typical variables such as crown diameter, tree height, trunk diameter, tree density and tree spacing. The emphasis is placed on the automatization of the process of identification of the most relevant image features for the forest structure retrieval task, exploiting both spectral and spatial information. Our approach is based on linear regressions between the forest structure variables to be estimated and various spectral and Haralick's texture features. The main drawback of this well-known texture representation is the underlying parameters which are extremely difficult to set due to the spatial complexity of the forest structure.

To tackle this major issue, an automated feature selection process is proposed which is based on statistical modeling, exploring a wide range of parameter values. It provides texture measures of diverse spatial parameters hence implicitly inducing a multi-scale texture analysis. A new feature selection technique, we called Random PRiF, is proposed. It relies on random sampling in feature space, carefully addresses the multicollinearity issue in multiple-linear regression while ensuring accurate prediction of forest variables. Our automated forest variable estimation scheme was tested on Quickbird and Pléiades panchromatic and multispectral images, acquired at different periods on the maritime pine stands of two sites in South-Western France. It outperforms two well-established variable subset selection techniques. It has been successfully applied to identify the best texture features in modeling the five considered forest structure variables. The RMSE of all predicted forest variables is improved by combining multispectral and panchromatic texture features, with various parameterizations, highlighting the potential of a multi-resolution approach for retrieving forest structure variables from VHR satellite images. Thus an average prediction error of  $\sim 1.1$  m is expected on crown diameter,  $\sim 0.9$  m on tree spacing,  $\sim 3$  m on height and  $\sim 0.06$  m on diameter at breast height.

© 2014 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

## 1. Introduction

Many studies have focused on estimating forest parameters from optical remote sensing data since the early years of satellite imagery. The methods consist usually in extracting one or several predictive variables for estimating tree or stand level variables that are useful for forest resources inventory and management such as tree height, stem diameter, density, basal area, stem volume or biomass (Kayitakire et al., 2006; Feng et al., 2010; Proisy et al., 2007;

Ozdemir and Karnieli, 2011; Song et al., 2010; Beguet et al., 2012). These studies are generally based on linear single regression modeling and only few multiple regression modeling. The accuracy in retrieving forest stand variables depends on the image spatial resolution. Hyypya et al. (2000) observed that medium or high resolution multispectral satellite imagery such as SPOT-4 (20 m resolution) or Landsat TM (30 m resolution) leads to a lower correlation performance, in comparison with sub-metric aerial photography. Using 5 m and 10 m SPOT-5 images, Wunderle et al. (2007), Wolter et al. (2009), Castillo et al. (2010) retrieved some forest stand attributes (such as crown diameter), exploiting image texture, with a good accuracy.

Over the last decade, a growing number of Very High Resolution (VHR) multispectral satellite images from various sensors has

\* Corresponding author at: EA 4592 G&E, ENSEGD, University of Bordeaux, 1 Allée F. Daguin, 33607 Pessac Cedex, France. Tel.: +33 5 57 12 10 36; fax: +33 5 57 12 10 01.

E-mail address: [benoit.beguet@ipb.fr](mailto:benoit.beguet@ipb.fr) (B. Beguet).

become available such as Pléiades, Quickbird, Geoeye, WorldView or Ikonos which provide sub-metric spatial resolution (0.5–1 m in panchromatic band, 2–4 m in multispectral bands).

VHR imagery provides a meaningful textural information. Various texture representations were proposed in literature and applied to the remote sensing data for a wide range of applications such as urban mapping (Soe and Tyler, 2004; Dell'Acqua and Gamba, 2003; Pacifici et al., 2009), image segmentation (Trias-Sanz et al., 2008; Gaetano et al., 2009) or vegetation structure mapping and habitat use (Wood et al., 2012; Tuttle et al., 2006).

Some recent studies (Kayitakire et al., 2006; Feng et al., 2010; Proisy et al., 2007; Ozdemir and Karnieli, 2011; Song et al., 2010; Beguet et al., 2012; Beguet et al., 2013; Gomez et al., 2012; Tuominen and Pekkarinen, 2005) have shown the potential of VHR imagery for forest inventory applications thanks to the strong relationship between forest spatial structure and image texture at stand level. Texture analysis of VHR satellite images generally applied to forest inventory can be divided into three main approaches (Maillard, 2003): wavelet or Fourier-based, variogram-based, and GLCM-based (Grey level Co-occurrence matrix) methods. Coueron et al. (2005), Proisy et al. (2007), Barbier et al. (2010) used a 2D Fourier periodogram to model canopy texture. Ruiz et al. (2004), Regniers et al. (2013), Van Coillie et al. (2007) used wavelets for forest structure analysis. A variogram is useful to explore the relationship between image texture and forest structure (Guyon and Riom, 1996; St-Onge and Cavayas, 1997; Franklin et al., 2001; Wolter et al., 1998; Song et al., 2010). Its main drawback is the requirement of a good non-linear model to fit the observed variogram before extracting texture indicators such as sill, range and nugget. The second order statistics derived from Grey Level Cooccurrence Matrix (GLCM) as defined by Haralick et al. (1973) are the most used texture features in forestry remote sensing literature (Franklin et al., 2001; Chehata et al., 2011; Boukir et al., 2013) and provide good performances for estimating forest parameters (Kayitakire et al., 2006; Wunderle et al., 2007; Wolter et al., 2009; Castillo et al., 2010; Ozdemir and Karnieli, 2011; Beguet et al., 2012). Generally, these parameters are set to fixed and hence unoptimal values. To tackle this major issue, an automated feature selection process that explores a wide range of texture parameter values is a worthwhile solution to investigate which, to the best of our knowledge, has not been proposed yet in literature. This automated spatial parameter tuning has an appealing property : it provides texture measures of diverse spatial parameters hence implicitly inducing a multi-scale texture analysis. It is well known that texture analysis is generally more relevant in a multi-scale context (Gonzales and Woods, 2008).

Moreover the existing studies generally involve only one spatial resolution, considering exclusively panchromatic or multispectral data. To our knowledge, only Wolter et al. (2009) have combined texture features from both kinds of data with different spatial resolutions, but via variograms and not GLCM.

Besides, the impact of image acquisition conditions on the forest structure retrieval from texture information is a critical issue. In fact, both view and sun angles influence the image texture, due to the interaction of radiation with vegetation structure. They particularly determine the length and the orientation of the shadows of crowns cast on ground, the fraction of sunny or shadowed crowns, and also the apparent radius or length of crown viewed by the sensor. Barbier et al. (2010) showed that the bidirectional variation in texture is to be accounted for when using a number of images with various angle configurations. In addition, image texture can significantly vary across the seasons due to phenological changes in vegetation structure, thus indicating that care is to be taken when measuring texture at different phenological stages (Culbert, 2009).

In this paper, we aim to fully exploit the potential of texture features extracted from VHR satellite images using a GLCM-based approach for estimating some typical forest stand variables, with a particular emphasis on automated parameter tuning, one of the major issues in texture analysis. The study focuses on the retrieval of the following forest structure variables: crown diameter, tree height and tree density or spacing which lead a spatial resolution-dependent image texture. The stem diameter at breast height is also considered, since it is easily measured on the field and it indirectly contributes to image texture via its correlation with the crown structure variables such as crown diameter. The top-tree height also contributes to image texture since it impacts on the length of crown shadows according to the solar elevation or the apparent crown length viewed by the sensor according to the view zenith angle.

The main objective of our work is to design and test an automated process for the selection of GLCM texture features and their optimal parameterization that is able to predict the forest variables with a satisfactory accuracy regardless of the season or the angle configuration of the remote sensing data. To achieve optimal results, numerous combinations of panchromatic and/or multispectral features were tested through multi-variable linear regressions involving field-measured forest variables. As collinearity is a very perturbing problem in multi-linear regression, this issue is carefully addressed through an innovative variable subset selection algorithm we called Random PRiF. A comparative analysis is carried out with other well-established variable subset selection techniques used in multi-variable regression, a classic dimensionality reduction: PCA (Principal Component Analysis) (Manly, 1994) and a recent one that is more adapted to prediction: LARS (Least Angle Regression) stepwise.

The developed automated forest variable retrieval scheme was assessed on stands of maritime pine (*Pinus pinaster* Ait.) of the forest of “Landes de Gascogne” in south-western France. An accurate estimation of the forest variables is aimed. Indeed, covering about one million hectares (75% of the total area of the region), these maritime pine stands are economically important at regional and national levels and thus an accurate inventory of wood resource is required. Two sites covering a large part of the structure diversity of this forest were considered, using VHR Quickbird and Pléiades images.

## 2. Material

### 2.1. Study sites

Both study sites are located in south-western France within the largest European maritime pine (*Pinus pinaster* Ait.) forest (Fig. 1), so-called *the forest of Landes de Gascogne*, which covers approximately one million hectares in a nearly flat area except for the coastal dunes. This forest consists of maritime pine even-aged stands which are intensively managed. The pine trees are planted in rows usually 4 m apart. The regeneration techniques have changed in recent years. Now the trees are generally regularly planted along the rows with a low density (~1000 trees/ha, i.e. within-row spacing ~2.5 m). Previously, the trees were always sowed with a higher density in the row (~5000 trees/ha, i.e. within-row spacing ~0.5 m). The trees are periodically thinned after clearing the understory vegetation. The clear-cutting occurs mostly when the pine trees are 50 years old.

The Nezer site, whose area is around 60 km<sup>2</sup>, is managed quite uniformly since it covers only two large tree farms. The stand size is large, the mean area is approximately 12 ha with a maximum of 50 ha. The stands are mostly rectangular and often circled by fire

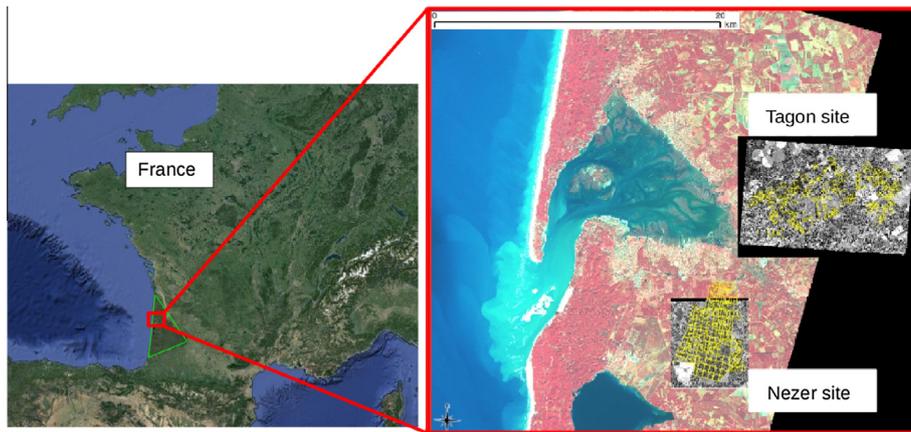


Fig. 1. Location of Nezer and Tagon study sites.

lanes or roads. The pine trees rows are oriented West–East in most of stands.

The Tagon site, which covers about 80 km<sup>2</sup>, is more complex and more heterogeneous than the Nezer one, since it is managed by many forest owners. The stands are smaller (mean area ~7 ha, maximum ~40 ha) and displays a larger variability in geometry, row orientation, and forest structure.

Both sites are highly representative of the whole maritime pine forest diversity in terms of forest structure, understory species composition, silvicultural and management practices.

## 2.2. Field data

For both sites, four forest structure variables were measured on field: diameter at breast height of trunk (*Dbh*), height of the tree top (*Ht*), crown diameter (*Cd*) and stand density (*Nah*). *Sp* is a tree spacing index that was calculated from *Nah* using a non linear function (cf. Eq. (1)) assuming each tree is located on the apex of an equilateral triangle. The field measurement campaigns have been driven by GIS using aerial photographs and the forest management data which were available. Thus several forest structure classes have been designed considering tree spatial organization and dimensions, in order to sample uniformly the whole range of forest variables variation. No measurements were performed on the very young stands (<4 year), where the tree cover fraction is very low and the undergrowth vegetation is often taller than the pine trees.

$$Sp(m) = \sqrt{\frac{20,000}{\sqrt{3} * Nah(\text{tree/ha})}} \quad (1)$$

For Nezer site, the measurements were made on 12 stands at the end of the year 2003. The trees sampled in each stand were located within a 80 m × 80 m square plot representative of the forest structure in the 200 m × 200 m area that encloses it.

For Tagon site, the measurements were made on 111 stands during summer 2012 at the end of the pine tree growth period. The trees sampled in each stand were included in a 20 m × 20 m square plot representative of the forest structure in the 80 m × 80 m area that encloses it.

The variation range of the forest variables (mean values per stand for *Cd*, *Ht*, *Dbh*) are summarized, for each site, in Table 1 and Fig. 2. Fig. 2 highlights the allometric relationships between variables *Cd*, *Sp* and *Ht*, i.e. their intrinsic correlations to the species and their correlation related to tree growth, site fertility and past silvicultural operations.

Table 1

Variation range of forest variables (Nezer: 12 sampled stands, Tagon: 111 sampled stands). Mean values per stand are given.

	<i>Cd</i> (m)	<i>Ht</i> (m)	<i>Nah</i> (tree/ha)	<i>Sp</i> (m)	<i>Dbh</i> (m)	<i>Age</i> (year)
<i>Nezer site</i>						
Min	2.95	9.3	189	3.04	0.15	13
Max	7.81	24.7	1253	7.81	0.46	51
<i>Tagon site</i>						
Min	0.77	1.7	150	1.31	0.02	4
Max	10.69	26.2	6729	8.77	0.56	68

## 2.3. Remote sensing data

VHR satellite images from two different sensors (Quickbird and Pléiades) were used. These images have a similar spatial resolution in similar spectral bands: ~0.5–0.6 m in panchromatic (Pan), and ~2–2.5 m in multispectral mode (MS). On Nezer site, one Quickbird image was processed and analysed. It was acquired on the 6th of October 2003. On Tagon site, two Pléiades images were acquired respectively on the 26th of June 2012 and the 8th of August 2012. All the images were geo-referenced and ortho-rectified. Their characteristics are given in Table 2.

This image set gives some repeated observations in different conditions of phenological stage of vegetation (summer/autumn), solar elevation (low/high) and angles of view (oblique/very oblique). The differences are higher between Nezer and Tagon datasets. For the Quickbird image acquired on Nezer in autumn, the understory vegetation was senescent with a low green leaf area whatever its species composition. Its contribution to the image reflectance spatial variation and consequently to image texture was probably low. On the contrary, the Pléiades images were acquired on Tagon in summer season when the green leaf area of the undergrowth vegetation is maximal. However, its impact on the reflectance can strongly vary in space according to the species (fern, heather, grass such as molinia and other) and their spectral properties.

## 3. Methods

The global framework of forest variable retrieval methodology is presented in Fig. 3. Test frames that are composed of image samples (reference stands) are constructed from both panchromatic and multispectral bands. Texture features with various parameterizations are then calculated on test frames. Median values are kept per image sample. These test frames are used for feature selection based on linear regression to keep the best texture features that are

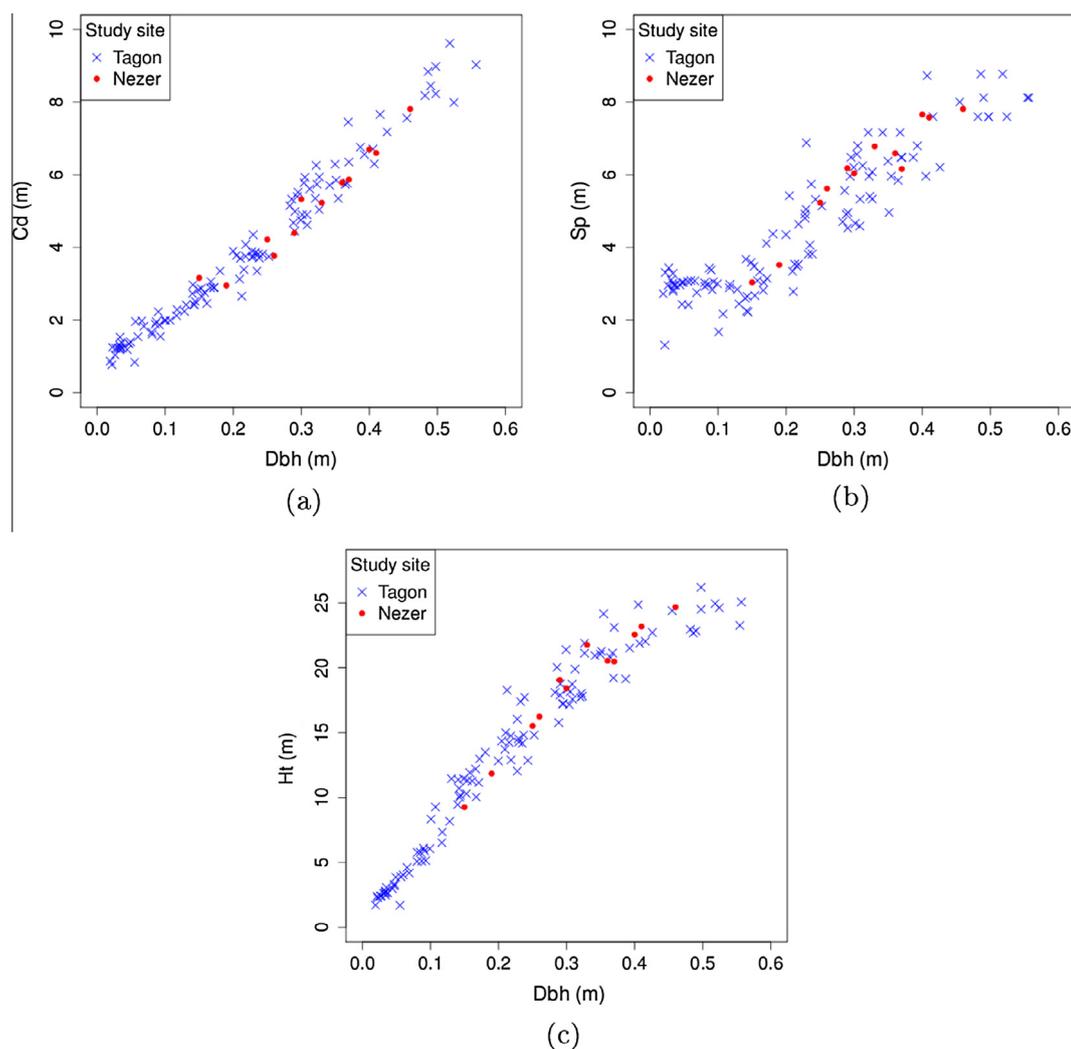


Fig. 2. Field measured forest variables on both sites: crown diameter (Cd), tree height (Ht) and tree spacing (Sp) as a function of diameter at breast height (Dbh).

**Table 2**  
Image acquisition parameters.

	QuickBird 06/10/2003	Pléiades 26/06/2012	Pléiades 08/08/2012
PAN-MS resolution (m)	0.62–2.48	0.5–2	0.5–2
Encoding (bits)	11	8	8
Panchromatic	450–900 nm	450–900 nm	450–900 nm
Blue	450–520 nm	430–550 nm	430–550 nm
Green	520–600 nm	500–620 nm	500–620 nm
Red	630–690 nm	590–710 nm	590–710 nm
Near Infra Red	760–900 nm	740–940 nm	740–940 nm
Sun azimuthal angle	157.3°	147.3°	154.3°
Sun elevation angle	33.1°	66.2°	59.3°
View azimuthal angle	42.5°	179.9°	180.0°
View zenithal angle	20.0°	24.8°	13.5°
Study site	Nezer	Tagon	Tagon

related to each forest variable. Both single and multiple linear regressions are processed. For multiple linear regression, a new variable subset selection technique, we called Random PRiF, is proposed. The whole process will be detailed here after.

### 3.1. Image features

In our approach, two first order texture features (mean and variance, called var) and eight second order GLCM texture features

were used (Haralick et al., 1973; Conners et al., 1984). The latter are considered as particularly relevant for forest applications (Franklin et al., 2001; Coburn and Roberts, 2004; Kayitakire et al., 2006; Murray et al., 2010; Wunderle et al., 2007; Ozdemir and Karnieli, 2011; Castillo et al., 2010; Chehata et al., 2014; Wood et al., 2012). They are described in Table 3.

All texture measures were calculated from panchromatic (Pan) and from four multispectral (MS) bands: Blue (B), Green (G), Red (R), Near Infra Red (NIR), with various combinations of GLCM parameter values. These parameters are: radius of the moving window  $r$  (window size =  $2r + 1$ ), displacement  $d$ , orientation  $o$  and quantification level number  $nbbin$ . A limited parameter range was fixed so as to make a trade-off between computational costs and effectiveness. The ranges of  $r$  and  $d$  were defined according to the spatial resolution and are given in Table 4. The quantification level  $nbbin$  was set to 8 as commonly used in remote sensing literature.

### 3.2. Test frames

Our texture feature selection process was automatically performed on test frames gathering small image samples of all stands where the forest variables were measured in field. This method made it easy to explore Haralick's texture features across a wide range of parameter values (GLCM parameters). Indeed thousands

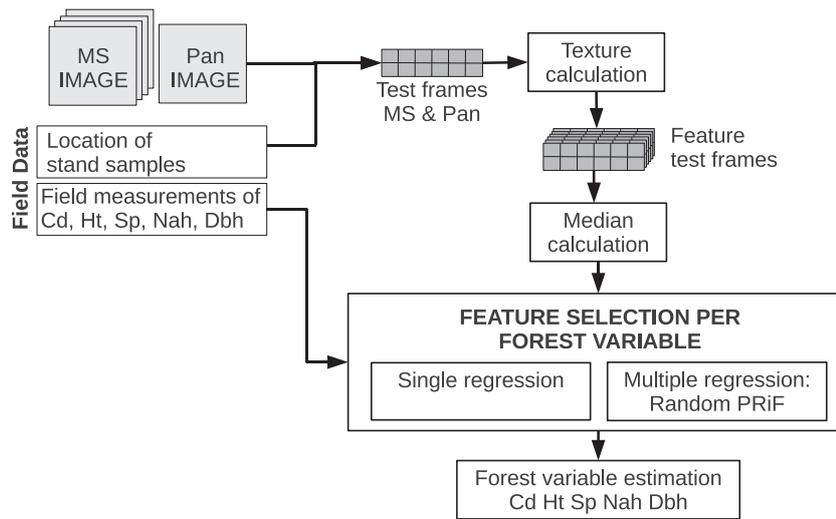


Fig. 3. A Global framework of forest variable retrieval.

Table 3  
GLCM texture features description.

Feature	Description	Formula
Energy (energy)	High when the GLCM is locally homogeneous	$\sum_{ij} g(i,j)^2$
Entropy (entropy)	Measure of the randomness of the intensity distribution	$-\sum_{ij} g(i,j) \log_2 g(i,j)$
Correlation (corr)	Linear dependency on neighboring pixels values	$\sum_{ij} \frac{(i-\mu)(j-\mu)g(i,j)}{\sigma^2}$
Inverse Difference (invdif)	A measure of homogeneous pixel values across an image, inversely related to Inertia	$\sum_{ij} \frac{1}{1+(i-j)^2} g(i,j)$
Inertia (inertia)	High values are indicative of high grey level values off the matrix diagonal	$\sum_{ij} (i-j)^2 g(i,j)$
Cluster Shade (clushade)	Higher order moment	$\sum_{ij} ((i-\mu) + (j-\mu))^3 g(i,j)$
Cluster Prominence (cluspro)	Higher order moment	$\sum_{ij} ((i-\mu) + (j-\mu))^4 g(i,j)$
Haralick's Correlation (harcorr)	Correlation measure	$\frac{\sum_{ij} (i,j)g(i,j) - \mu_i^2}{\sigma_i^2}$

with:

- $g(i,j)$  a normalized GLCM matrix element.
- $\mu = \sum_{ij} i \cdot g(i,j) = \sum_{ij} j \cdot g(i,j)$  is the pixel values weighted mean.
- $\sigma = \sum_{ij} (i - \mu)^2 \cdot g(i,j) = \sum_{ij} (j - \mu)^2 \cdot g(i,j)$  (due to matrix symmetry) is the pixel values weighted variance.

Table 4  
Values of the spatial GLCM parameters.

Data	Radius $r$ (pixel)	Displacement $d$ (pixel)	Orientation $\theta$ (degree)
Pan	5–25 step: 5	1–10 step :1	0–135° step: 45°
MS	3–12 step: 3	1–4 step:1	0–135° step: 45°

of descriptors were calculated given the large number of possible combinations of parameter values. The proposed feature selection scheme assesses the capability of diverse multiple features descriptors to retrieve the forest structure with minimal computational costs. Fig. 4 exhibits the few image samples used to build the test frames. Each image sample covers a square area centered at a field measurement plot (cf. Section 2.2) with a slightly larger size. Its width is 200 pixels in Pan and 50 pixels in MS (i.e. ~120 m) on Nezer site. It is 120 pixels in Pan and 30 pixels in MS for Tagon site (i.e. ~60 m).

The maximal and minimal values of radiometry observed on the whole test frame were used to define the GLCM quantification levels in each spectral band of each image in order to guarantee a radiometry resolution for forest pixels in concordance with the used number of levels (nbbin = 8). For each sampled plot, the

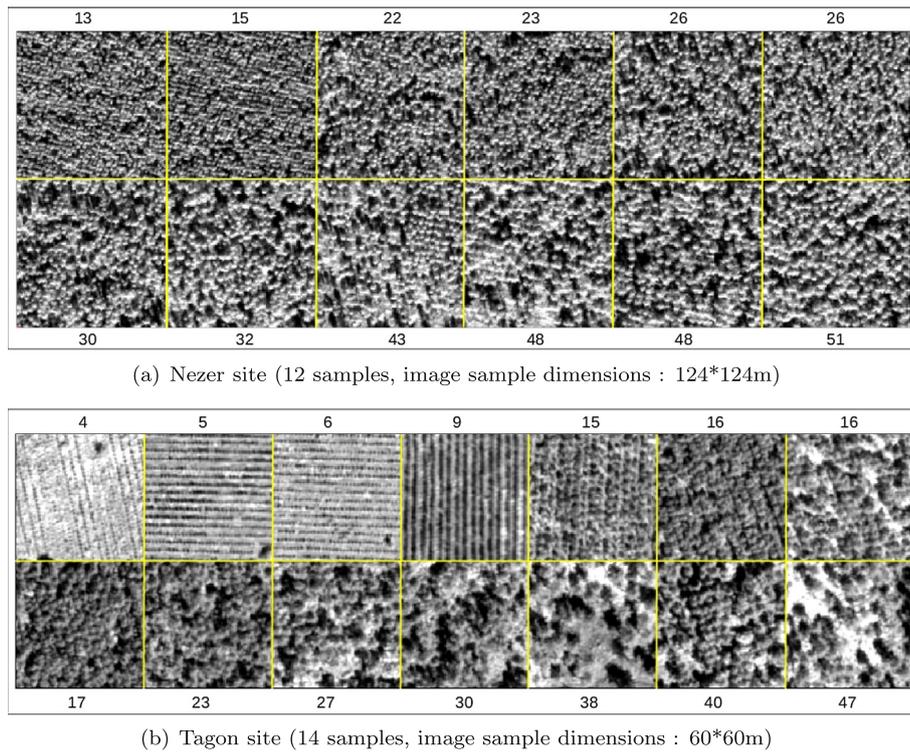
median value of each tested image feature was calculated and used as a sample centroid in the next step of selecting the most relevant features for retrieving the forest variables. The median value was preferred to the mean value, which is commonly used in remote sensing for this matter, in order to overcome smoothing effects.

### 3.3. Automated feature selection using linear regression

The selection of the most relevant texture features for retrieving each of the five forest variables is based on statistical modeling of forest variable as linear function of one or multiple image features. Modeling is based on linear regression. Single linear regression assumes that only one texture feature is sufficient to estimate the considered forest variable and thus exploits only one spectral band and only one spatial resolution. Multiple linear regression is applied on both Pan and MS texture features and thus allows to assess a multi-scale and multi-resolution modeling. Thus better accuracies are expected but multicollinearity effects have to be taken into account carefully.

#### 3.3.1. Single variable linear regression

The linear regressions were performed between all forest structure variables and all texture features, derived from Pan and MS



**Fig. 4.** Subset of image samples in panchromatic band for both sites illustrating the variability in forest structure of the sampled stands. The samples are sorted by increasing tree age.

bands with different parameterizations. The determination coefficient  $R^2$  was used as criterion to identify the best texture feature in Pan and in MS datasets for each date and each site. The processing was automated to select the texture feature leading to the highest  $R^2$  with the forest variable. The accuracy of the linear modeling was estimated with the RMSE (root mean square error) calculated by the leave-one-out cross validation technique (named here LOO RMSE).

### 3.3.2. Multiple variable linear regression

Multiple linear regression has been applied in literature using SPOT-5 multispectral information only (initial bands, derived vegetation indexes and GLCM texture bands) (Castillo et al., 2010). Wunderle et al. (2007) used image texture derived from pan-sharpened SPOT imagery and stepwise regression analysis to model a stand structural index. Ozdemir and Karnieli (2011) modeled forest structure parameters as function of WorldView2 spectral texture measures using stepwise multi-linear regressions. Our approach explores the potential of combining Haralick's features derived from both Pan and MS bands, thus using jointly different spatial and spectral resolutions. In addition, the underlying automated processing allows to optimize GLCM's spatial parameters while exploring various texture feature combinations thus allowing multi-scale modeling. The multiple regressions have to be thus performed on thousands of descriptors, many of which being likely correlated. The main difficulty in multiple linear regression is minimizing the multicollinearity of the feature subset solutions in order to generate stable models and avoid over-fitting. This is generally done using the Variance Inflation Factor (VIF), a good multicollinearity detector which is defined as follows for any variable  $j$  used as regressor in multiple regression:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (2)$$

where  $R_j^2$  is the determination coefficient obtained by regressing the  $j$ th regressor on the remaining regressors. A variance inflation factor exists for each of the  $j$  regressors in a multiple regression model. The higher  $VIF_j$ , the higher the collinearity between variable  $j$  and the other variables. In other words,  $VIF_j$  equals to 1 if there is no correlation between them. VIF values exceeding 4 warrant further investigation, while VIF values exceeding 10 are signs of serious multicollinearity requiring correction (Castillo et al., 2010; Haan, 2002; Kutner et al., 2004).

A common prediction criterion involved in multilinear regression is *PRESS* (Predicted Residual Error Square Sum) statistic (Allen, 1974) defined as:

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i(-i))^2 \quad (3)$$

where  $\hat{y}_i(-i)$  is the prediction for observation  $i$  using the model calculated from the  $n - 1$  others.

*PRESS* is equivalent to a statistic of Leave-One-Out cross validation (LOO RMSE =  $\sqrt{PRESS/n}$ ).

Various methods for feature subset selection exist using multiple regressions. In previous work (Beguet et al., 2012), the performances of three methods were compared: the classic step-by-step forward method, a LASSO (Least Absolute Shrinkage and Selection Operator) method, and a forward stepwise approach based on LARS (Least Angular Regression) (Efron et al., 2004). The latter is a newer forward stepwise method. Unlike the classic Forward method which incrementally adds the variable leading to the highest F-statistic, the LARS algorithm adds step-by-step the variable that better explains the current residuals. LARS stepwise turned out the most effective subset selection method, with both low *PRESS* and low *VIF* (below the critical value of 4 for each regressor). The main drawbacks of these step-by-step variable selection techniques are (1) they provide suboptimal solutions due to the underlying incremental optimization scheme and (2) the relative importance of

the selected variables are highly imbalanced and biased towards the first variable.

To alleviate the weakness of traditional step-by-step variable selection, we propose a new variable subset selection scheme, that we called “Random PRiF”. Unlike LARS stepwise which is an incremental method (keeping the first best feature and adding some complementary ones) and hence suboptimal, our method is global and based on a simple yet statistically robust optimization scheme: random sampling (with replacement) in feature space. The main idea is to assess, through an iterative process, a large number of possible subsets of image features chosen randomly (with replacement) among the whole range of descriptors, by minimizing a new statistical criterion which appropriately combines both *PRESS* and *VIF* criteria. The aim is to automatically find a feature subset with a strong predictive performance while keeping a low *VIF*. The criterion we named *PRiF* is defined as follows:

$$PRiF = PRESS * \frac{\exp(V)}{V^2} \quad (4)$$

where *V* is the maximum of the *m*-component *VIF* vector associated to the *m* selected features.

The use of a function of *VIF* in *PRiF* that has a hyperbolic behavior favors feature subsets with a good prediction performance while penalizing those with high values of *VIF* in order to achieve both accuracy and stability in resulting models. Our technique provides optimal variable subset solutions whose relative importances are significantly more balanced than their step-by-step solutions counterpart.

Though inducing significantly higher computational costs than step-by-step approaches, the latter faces much higher risks of falling into local optima, leading to poor models. In Random PRiF, the higher the number of iterations of the random sampling-based procedure, the higher the achievable prediction accuracy and the lower the risk of a suboptimal solution. A comparative analysis is conducted in the following section between Random PRiF and two well-established methods: PCA (Principal Component Analysis) dimensionality reduction method and LARS stepwise. PCA has been chosen as a reference method without multicollinearity, i.e. *VIF* = 1 between the principal components used as regressors. LARS stepwise, a recent state-of-the-art variable subset selection method, is considered for performance comparison because it led to the best results in a previous comparative analysis (Beguet et al., 2012).

## 4. Results and discussion

The image features calculation was processed with the Orfeo-ToolBox library (OTB) <http://www.orfeo-toolbox.org/otb/>, the statistical analysis was carried out on R software. The whole processing chain was managed using Python programming language.

### 4.1. Single feature solutions

The single regressions were processed separately on Pan and MS datasets for both sites. Due to the diversity of the tested parameter values, 1730 and 2368 texture feature descriptors were tested for Pan and MS respectively. For each forest variable, numerous significant linear relationships with image features were found (considering a *p*-value <0.01 for *R*<sup>2</sup>). Features with the highest *R*<sup>2</sup> are presented in Table 5 for Pan and MS datasets on Nezer and Tagon sites respectively. For each forest variable, the best PAN and MS features are presented with the corresponding parameterization and spectral band.

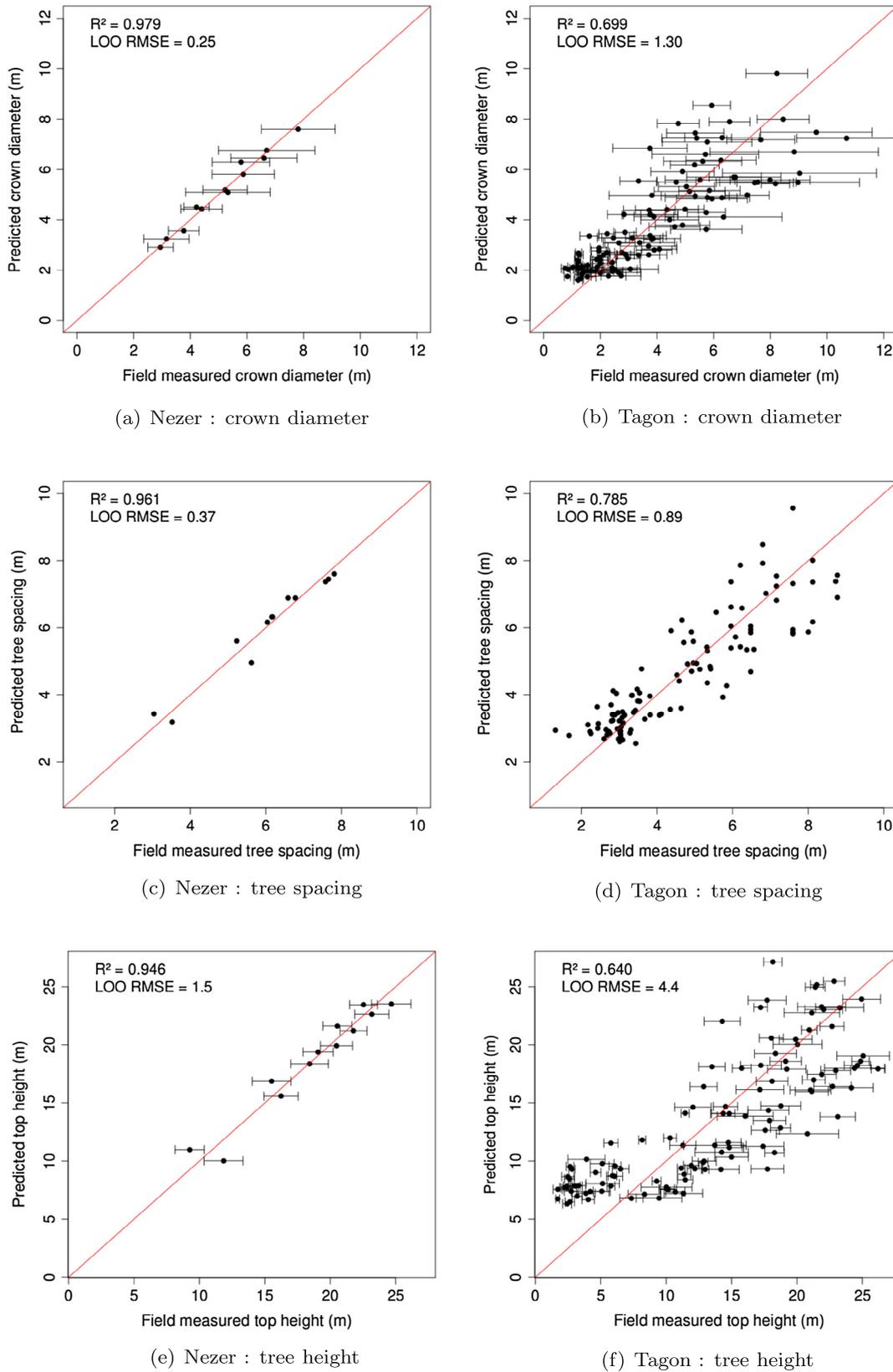
For Nezer site (Table 5), all considered forest variables were well retrieved by image texture features (*R*<sup>2</sup> varies from 0.89 to 0.98). The performances are mostly similar for Pan and MS on this reduced set of stands (*n* = 12). The texture features calculated on the Green band appeared to be the most effective MS ones. Second order texture features were clearly the most descriptive of the forest structure. In this case, inertia was the most frequently selected texture feature, for both resolutions and with various parameterizations. It is interesting to notice that for Pan features, window radius parameter varied from 5 to 25 pixels depending on the retrieved forest variable. This confirms that no a priori presumption could be made on GLCM parameters, thus justifying a wide-ranging and automatic search of their optimal values for the forest variable retrieval task.

For Tagon site (Table 5), the single regressions were processed for both dates: June and August. However due to paper length restriction, only June results are presented. The quality of predictions is lower than the one observed for Nezer. *R*<sup>2</sup> varies from 0.32 (Nah) to 0.73 (Sp) in June and from 0.34 (Nah) to 0.78 (Sp) in August and LOO RMSE is always higher. Similarly to Nezer site, the Green band provides the best single MS-based models. For both dates over Tagon, MS features appeared to be more relevant for forest structure estimation than PAN ones. Besides, the variance, a first order feature, gives some best results. Many factors such as acquisition conditions, sun position, phenology stage and spectral properties of the understory may influence this result. For instance, since Tagon site is characterized by a high variety of tree row orientations, the variance, which is insensitive to image texture anisotropy, could appear more robust to orientation variations than second order features. Similar values for *R*<sup>2</sup> and RMSE were achieved on June and August images, though slightly better for the August image, but with different best performing image features. Indeed, the image texture of the same stands varies probably between both dates due to small differences in phenology, sun and view angles. For instance, the spatial heterogeneity of the understory reflectance was probably slightly smaller in August due to the phenological stage (the leaf growth ended in August for all species unlike in June), the solar position induced slightly larger shadows on ground in August, also a better visibility of the tree dimensions

**Table 5**

The best performing single MS and Pan image features for retrieving each forest variable. Nezer site, October image. Tagon site, June image.

Forest var.	Band	Feature	<i>r</i>	<i>d</i>	<i>o</i>	<i>R</i> <sup>2</sup>	<i>p</i> Value	LOO RMSE
<i>NEZER</i>								
Cd	B	cluspro	3	1	0	0.970	<10e−06	0.30
	Pan	inertia	10	5	45	0.979	<10e−06	0.25
Sp	G	inertia	3	2	0	0.961	<10e−06	0.37
	Pan	inertia	5	1	45	0.944	<10e−06	0.40
Ht	G	inertia	6	2	0	0.946	<10e−06	1.5
	Pan	inertia	25	1	45	0.905	<10e−06	1.6
Nah	B	corr	3	1	45	0.891	<10e−06	140
	Pan	inertia	5	1	135	0.909	<10e−06	37
Dbh	G	inertia	12	1	45	0.938	<10e−06	0.029
	Pan	inertia	10	5	45	0.927	<10e−06	0.029
<i>TAGON</i>								
Cd	G	var	6			0.650	<10e−06	1.397
	Pan	corr	10	2	45	0.505	<10e−06	1.659
Sp	G	var	3			0.737	<10e−06	0.979
	Pan	var	25			0.572	<10e−06	1.242
Ht	NIR	invdif	12	3	135	0.635	<10e−06	4.450
	Pan	corr	15	3	45	0.550	<10e−06	4.993
Nah	NIR	corr	12	1	135	0.445	<10e−06	682.112
	Pan	corr	10	2	45	0.321	<10e−06	716.776
Dbh	G	var	6			0.650	<10e−06	0.086
	Pan	corr	10	2	45	0.533	<10e−06	0.099



**Fig. 5.** Field measured forest variables vs. image derived estimates from best single variable models. Left: Nezer site, October image. Right: Tagon site, August image. Bars correspond to  $\pm 1$  standard deviation.

was probably obtained with a lower view zenith angle in August. In spite of these small differences, the image texture was very informative on the forest structure for both dates. The

performances of the best image feature subsets are relatively good and similar for both dates, thus seeming to be rather robust to image acquisition conditions.

For both sites and dates, the best prediction performances were obtained for tree spacing (*Sp*), crown diameter (*Cd*) and tree height (*Ht*) which are the forest variables that directly impact the image texture. However the trunk diameter *Dbh*, which has no direct effect on texture, also showed high  $R^2$  values close to those obtained for *Cd*. This comes from the strong linear allometric relationships between both forest variables (see Fig. 2) already demonstrated for maritime pine (e.g. Porté et al. (2000)). The tree density *Nah* showed the lowest linear correlations with texture features unlike *Sp* which is calculated as a non-linear transform of *Nah* (cf Eq. (1)).

Fig. 5 compares the values measured on field and their estimates predicted from the best single variable regression model for *Cd*, *Sp* and *Ht*. No outliers are noticeable on residuals. The presumption of linear relationships for the three forest variables are clearly confirmed for the Nezer site, but the small number of observations does not allow a reliable interpretation of the results. The Tagon site is more informative since it provides a higher structure variability of maritime pine forest stands.

First, the variance (see standard deviations in Fig. 5) of field measured variables has to be considered when interpreting the quality of predictions. A heteroscedastic behavior is observed, more pronounced for crown diameter *Cd* than for tree height *Ht*. This highlights the fact that the older the stand, the higher the intra-stand variance of *Cd*. Indeed, with an increasing age and spacing between trees, the tree spatial distribution becomes more heterogeneous, thus inducing local variations in crown growth (due to tree competition) and thus in *Cd*. This also points out the uncertainty on *Cd* field measures. The higher the trees, the more difficult to measure the crown diameter.

Secondly, the relationships between measured and predicted values are not clearly linear except for *Sp* (Fig. 5(d)). The predicted values for *Cd* and *Ht* were overestimated for the youngest stands (Fig. 5(b) and (f)). For instance, for field measured *Ht* < 5 m, the predicted values varies from 5 to 10 m. Moreover, for the three forest variables, the residual variance is proportional to field measured variable values. Thus using a single regression model is not really suitable for estimating a forest structure variables.

Finally these preliminary results showed the potential of an automated selection of the most informative texture descriptors (i.e. the best Haralick feature parameterization) to estimate forest structure variables.

#### 4.2. Multiple feature solutions

For Nezer site, the number of texture features per regression was fixed to 3 using Mallow's  $C_p$  curve (Beguet et al., 2012). It was set to 5 for Tagon, as a trade-off between parsimony and complementarity of information by various texture features. The maximal iteration number used to obtain the best solution was set experimentally to 1 million for both sites.

##### 4.2.1. Performance analysis of feature selection methods

The three competing feature selection methods (PCA dimensionality reduction, LARS stepwise and Random PRiF) were compared, on both sites, for all forest variables (5) and using three image datasets per date: Pan alone (texture calculated on Panchromatic band), MS alone (texture calculated on multispectral bands), and MS-Pan which is composed of all features from both resolutions (hereinafter called multi-scale MS-Pan dataset). Hence, 30 different experiments (2 sites \* 3 image dataset types \* 5 forest variables) comparing the three methods were repeated. For Tagon, only August image data was tested. Table 6 illustrate some of the results by showing the performances ( $R^2$ , PRESS, VIF, LOO RMSE) of the best performing feature subsets obtained from each of the three methods on the MS-Pan dataset of Nezer site and on the Pan dataset of Tagon site.

A voting procedure was applied on the 30 experiments to identify which feature selection method was the most relevant to predict forest variables using as criterion the smallest LOO RMSE. Our Random PRiF feature selection method got the best overall prediction performance in almost 77% of cases, followed by the LARS stepwise. The PCA reduction is the worst method in all cases (cf. Table 6). These results confirm the relevance and the effectiveness of the introduced PRiF criterion within a random global search context. In addition, the collinearity, which is a critical issue in multi-linear regression, was low with maximum VIF values lower than the established critical threshold of 4 (Castillo et al., 2010) for all best solutions of both LARS stepwise and Random PRiF (cf. Table 6). As for PCA, which builds a new feature space composed of uncorrelated variables (of decreasing importance) and hence ensures by definition no collinearity, VIF is not a relevant criterion (VIF = 1 for all variables of a given subset). The results show that our Random PRiF method has a good behavior in terms of model stability and reliability, exhibiting maximal VIF values lower than 4 in all cases. LARS stepwise generally outperforms Random PRiF with respect to VIF criteria while it is the opposite with respect to prediction quality criteria (PRESS and LOO RMSE). But, this is the very essence of our novel variable selection criterion which was built so as to increase prediction accuracy while keeping collinearity reasonably low. We deliberately favors PRESS over VIF criterion because the former has a priori more impact on multi-linear overall model performances than the latter. PCA obviously provides the lowest possible VIF values but this never allowed this well-known method to outperform the two other assessed methods in all the 30 forest variable retrieval carried out experiments.

Besides, our Random PRiF method provided more balanced texture feature subsets than LARS stepwise with regard to the "variable importance" statistics. Variable importance is defined as the contribution of each regressor in the obtained model ; it was estimated through LMG statistic (Gromping, 2006). This is another appealing feature of our method that was also designed to alleviate some weaknesses of traditional step-by-step variable subset selection methods such as LARS stepwise (Efron et al., 2004) which provide suboptimal and highly imbalanced (large bias towards the first selected feature) solutions. Our variable subset selection methods relies on a simple and statistically robust optimization scheme: random sampling in feature space.

Table 7 outlines the importance of each texture feature in the selected subsets by both LARS stepwise and RandomPRiF methods for Nezer site. Random PRiF method led to significantly lower maximal importance values than LARS stepwise which exhibits a very large bias towards the first feature. For instance, for *Ht* variable, Random PRiF provides a balanced subset where the three texture features are roughly of same importance while LARS stepwise leads to a completely unbalanced feature subset. Indeed, the latter method is based on an incremental optimization scheme which is not the optimal way to combine features in a multiple regression framework. Though leading to suboptimal solutions, this step-by-step method has an advantage over our method : its lower computational complexity. This balanced behavior of Random PRiF method was clearly observed on Nezer site as well as on Tagon site though less dramatically. Thus the random global optimization approach used in Random PRiF helps to optimize the complementarity between features. It also leads to higher diversity in texture information capturing the essence of forest structure as opposed to step-by-step methods which more or less tend to reduce the potential of multiple-feature models to single-feature ones.

Finally, the results of the different experiments previously mentioned were summarized per forest variable and per site, a sample of which composed of the best performing pair (dataset/feature selection method) is presented in Table 8. It can be noticed that, the multi-scale MS-Pan dataset generally leads to the best models.

**Table 6**

Feature selection methods performance comparative analysis : statistics criteria of best feature subset using the MS-Pan dataset. October image on Nezer site (3 features subset), August image on Tagon site (5 features subset).

Forest var.	Selection method	multiple $R^2$	PRESS	Max(VIF)	LOO RMSE
<i>NEZER</i>					
Cd	PCA	0.917	5.458	1.000	0.67
	LARS stepwise	0.995	0.190	1.288	0.13
	Random PRiF	0.996	0.168	2.421	0.12
Sp	PCA	0.906	6.149	1.000	0.72
	LARS stepwise	0.994	0.535	1.025	0.21
	Random PRiF	0.996	0.233	2.734	0.14
Ht	PCA	0.903	66.034	1.000	2.3
	LARS stepwise	0.992	5.114	1.070	0.7
	Random PRiF	0.994	2.791	2.519	0.5
Nah	PCA	0.871	464568	1.000	197
	LARS stepwise	0.998	8154	1.287	26
	Random PRiF	0.997	5582	3.404	22
Dbh	PCA	0.892	0.029	1.000	0.049
	LARS stepwise	0.987	0.003	1.226	0.016
	Random PRiF	0.993	0.001	1.270	0.009
<i>TAGON</i>					
Cd	PCA	0.534	331.114	1.000	1.73
	LARS stepwise	0.752	180.304	1.357	1.27
	Random PRiF	0.800	129.015	2.293	1.08
Sp	PCA	0.548	214.820	1.000	1.39
	LARS stepwise	0.765	116.219	2.001	1.02
	Random PRiF	0.808	84.263	1.860	0.87
Ht	PCA	0.556	2904.690	1.000	5.1
	LARS stepwise	0.837	1144.369	2.052	3.2
	Random PRiF	0.851	1032.486	1.666	3.1
Nah	PCA	0.225	66281517	1.000	773
	LARS stepwise	0.523	51588535	1.408	682
	Random PRiF	0.500	46197877	1.748	645
Dbh	PCA	0.541	1231.149	1.000	0.105
	LARS stepwise	0.818	475.571	3.817	0.065
	Random PRiF	0.823	461.019	2.132	0.064

**Table 7**

Variable importance in the subsets of 3 features selected by LARS stepwise and Random PRiF methods, Nezer site, MS-Pan dataset.

Forest var.	Texture variable importance	
	LARS stepwise	Random PRiF
Cd	0.921	0.752
	0.011	0.272
	0.066	0.220
Sp	0.959	0.445
	0.035	0.442
	0.004	0.111
Ht	0.975	0.372
	0.018	0.326
	0.006	0.300

**Table 8**

Best multiple-variable solutions on Nezer and Tagon (August) sites with respect to both feature selection method and dataset type.

Forest var.	Site	Dataset	Feature selection method
Cd	Nezer	MS-Pan	Random PRiF
	Tagon	Pan	Random PRiF
Sp	Nezer	Pan	Random PRiF
	Tagon	MS-Pan	LARS Stepwise
Ht	Nezer	MS	Random PRiF
	Tagon	MS-Pan	LARS Stepwise
Nah	Nezer	MS-Pan	Random PRiF
	Tagon	MS-Pan	Random PRiF
Dbh	Nezer	MS-Pan	Random PRiF
	Tagon	MS-Pan	LARS Stepwise

This highlights the potential of combining both spatial and spectral resolutions. Regarding the performances of feature selection methods over 30 experiments, Random PRiF appears as the best alternative. As a result from these findings, the following sections focus only on the solutions provided by Random PRiF feature selection method using MS-Pan datasets.

4.2.2. Contribution of multi-scale texture features

The best performing texture feature subsets obtained from MS-Pan dataset using random PRiF selection method are presented in Table 9 for Nezer and Tagon (only August image) sites, respectively. For both sites, the selected subsets combined in most cases both Pan and MS features with various parameterizations in

window size ( $r$ ), displacement length ( $d$ ) and orientation ( $\theta$ ). It can be noticed that the multi-linear regression combines various second order texture features with different parameterizations and thus overcomes the anisotropy issue (i.e. tree spatial distribution by rows). For instance, the best feature subset to retrieve Nah on Nezer site combines two resolutions and three different orientations (45°, 90° and 135°).

Another example is the best feature subset to retrieve Cd on Tagon site which has a higher forest structure variability than Nezer site. It is composed of two GLCM moments in Pan with the same window size (15 pixels) and different displacements (1 and 5 pixels) providing multi-scale analysis. In addition, three lower

**Table 9**

Multiple solutions combining panchromatic and multispectral data by Random PRiF, Nezer site with October image and Tagon site with August image.

Forest var.	Band	Feature	<i>r</i>	<i>d</i>	<i>o</i>	multiple $R^2$	VIF	LOO RMSE
<i>NEZER</i>								
Cd	B	cluspro	3	1	0	0.996	2.421	0.12
	Pan	corr	25	9	135		2.382	
	B	corr	9	2	0		1.039	
Sp	Pan	invdif	5	2	135	0.996	2.735	0.14
	Pan	clushade	20	7	135		1.998	
	G	cluspro	6	4	0		2.412	
Ht	R	corr	9	3	0	0.994	1.164	0.5
	R	cluspro	6	2	0		2.346	
	R	inertia	3	1	90		2.519	
Nah	NIR	corr	12	2	45	0.997	1.069	22
	R	energy	3	2	90		3.404	
	Pan	inertia	15	2	135		3.389	
Dbh	Pan	invdif	10	4	45	0.993	1.270	0.009
	R	corr	3	2	0		1.062	
	G	entropy	6	3	135		1.282	
<i>TAGON</i>								
Cd	Pan	harcorr	15	1	135	0.807	2.261	1.10
	Pan	cluspro	15	5	0		2.240	
	NIR	harcorr	12	5	0		2.140	
	NIR	clushade	3	2	45		1.009	
	NIR	inertia	12	3	135		2.128	
Sp	Pan	harcorr	20	8	0	0.819	1.127	0.84
	B	harcorr	9	2	0		1.918	
	R	cluspro	12	1	45		2.23	
	R	var	6				2.303	
	G	harcorr	3	1	45		2.100	
Ht	R	cluspro	9	1	0	0.831	1.607	3.1
	Pan	entropy	25	4	0		1.778	
	Pan	inertia	25	1	0		1.256	
	Pan	clushade	15	10	135		1.148	
	Pan	inertia	25	2	90		1.290	
Nah	NIR	harcorr	9	2	0	0.625	1.562	586
	G	clushade	3	1	0		1.107	
	G	inertia	6	5	0		2.106	
	NIR	corr	3	1	135		1.225	
	NIR	invdif	12	5	0		2.526	
Dbh	Pan	inertia	20	2	0	0.808	1.055	0.066
	G	var	6				1.192	
	G	corr	12	1	135		1.188	
	B	harcorr	12	1	90		2.021	
	G	harcorr	9	5	0		2.070	

resolution texture features, calculated on the NIR band, with different parameterizations, completed the subset.

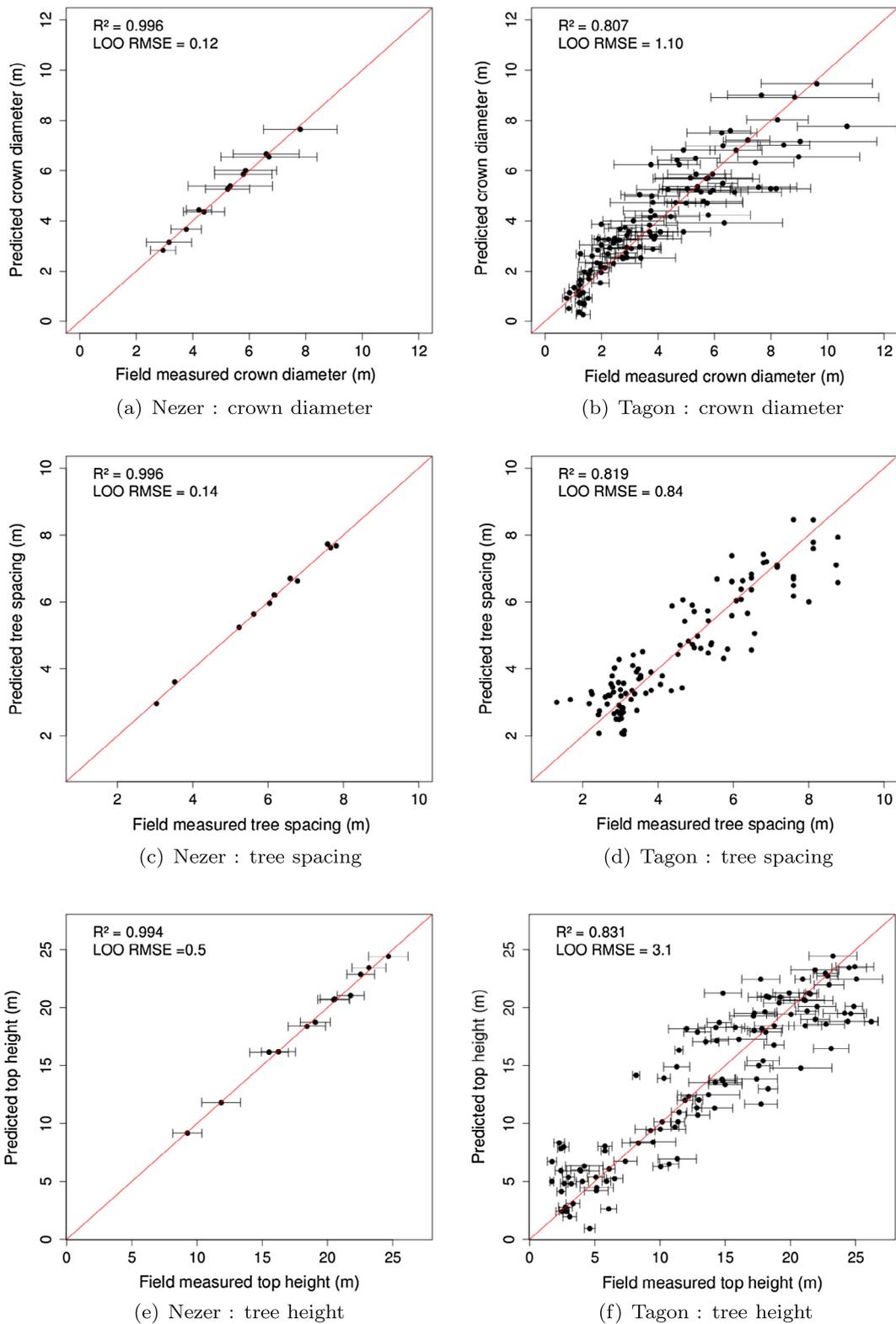
While single feature solutions (*cf.* Section 4.1) were essentially first order variance features for Tagon site, which are insensitive to row direction variability, multiple feature solutions took advantage of the GLCM texture features combining them with different parameterizations automatically. Moreover, as previously observed for single solutions, the obtained multiple solutions, for both dates on Tagon site, are different but their prediction performances ( $R^2$ , LOO RMSE) are similar. Thus the proposed automated selection of best texture features allows to exploit the texture information related to the considered forest variable regardless of tree row direction variability and image acquisition conditions. Fig. 6 gives a comparison between the values measured in field and their estimates predicted from the best multiple variable regression models for *Cd*, *Sp* and *Ht*. The multiple variable models clearly reduce the residuals in comparison with single variable models (*cf.* Fig. 5) for both sites. The residuals are also distributed quite randomly around the line 1:1; reflecting a nearly constant residual variance whatever the value of the forest variable unlike single solutions. As a result, the individual prediction error was minimized with no noticeable over or underestimation over the variation range of each forest variable in comparison with the

uncertainties on their values measured in field (see the length of bar indicating standard deviations).

In Table 10, the overall performances of single and multiple solutions in retrieving forest variables are compared according to LOO RMSE for both sites based on all available images (one date for Nezer and two dates for Tagon).

All prediction scores are improved with multiple solutions. On Nezer site, LOO RMSE was strongly improved for *Cd* by 52%, for *Ht* by 66% and for *Sp* by 62%. On Tagon site (August), which has a larger range of variation in forest variables and a larger forest structure variability between and within sampled stands compared to Nezer site, LOO RMSE was improved by 15.3% for *Cd*, by 29.5% for *Ht* and by 6.7% for *Sp*.

The gain in accuracy is better for *Dbh* (69%, 28%, 20% for Nezer, Tagon in June and in August, respectively) than for *Cd* (52%, 20%, 15%). Even if *Cd* and *Dbh* are highly correlated (*cf.* Fig. 2)) and *Cd* impacts directly the image texture unlike *Dbh*, this could be explained by a higher relative uncertainty in field measurement of *Cd* (see standard deviations in Fig. 6(a) and (b)). The gain is high for *Ht* (67%, 20%, 30% for Nezer, Tagon in June and in August, respectively), due probably to the greater complexity of the contribution of this variable to texture which is related to the length and direction of the crowns and the corresponding shadows as viewed



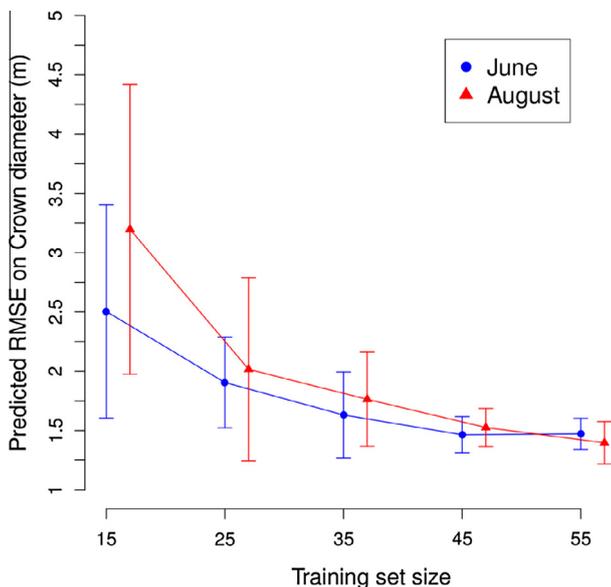
**Fig. 6.** Field measured forest variables vs. image derived estimates from best multiple variable models provided by Random PRiF. Left: Nezer site, October image. Right: Tagon site, August image. Bars correspond to  $\pm 1$  standard deviation.

by the sensor. The gain is the lowest for  $Sp$  in Tagon (6% in June and 7% in August). The single solution accuracy was already high for Nezer with a limited number of sampled plots ( $n = 12$ ) and the forest structure is more homogeneous within the sampled stands

with a lower range of variation between them (age > 12 years). These conditions were optimal to produce a quasi-linear mono-variable model, which is likely to be unrobust for prediction over other stands.

**Table 10**  
Performance comparison of single and multiple solutions obtained on the two sites from the MS-PAN dataset.

Forest var.	Site	LOO RMSE		Variable range
		Single	Multiple	
Cd (m)	Nezer	0.25	0.12	2.95–7.81
	Tagon (June)	1.40	1.12	0.77–10.69
	Tagon (August)	1.30	1.10	0.77–10.69
Sp (m)	Nezer	0.37	0.14	3.04–7.81
	Tagon (June)	0.98	0.92	1.31–8.77
	Tagon (August)	0.89	0.83	1.31–8.77
Ht (m)	Nezer	1.5	0.5	9.3–24.7
	Tagon (June)	4.5	3.6	1.7–26.2
	Tagon (August)	4.4	3.1	1.7–26.2
Nah (tree/ha)	Nezer	37	22	189–1253
	Tagon (June)	682	640	150–6729
	Tagon (August)	677	586	150–6729
Dbh (m)	Nezer	0.029	0.009	0.15–0.46
	Tagon (June)	0.086	0.062	0.02–0.56
	Tagon (August)	0.081	0.065	0.02–0.56



**Fig. 7.** Average RMSE distribution as function of the number of samples in training set. Results for Cd from same test set ( $n = 40$ ) and with 30 iterations per different size of the training set. Bar =  $\pm 1$  standard deviation.

Finally, the best LOO RMSE observed for Tagon using multiple solutions are very satisfactory for Cd and Sp, since they are very close to the expected accuracy with respect to image resolution: LOO RMSE  $\sim 1.1$  m for Cd and  $\sim 0.8$  m for Sp, i.e.  $\sim$  twice the resolution in Pan (0.5 m). Thus, for the youngest stands with tree size or spacing smaller than 1 m, it would be hard to find a model providing a good accuracy.

These results show that the multiple solutions are suitable to produce forest variable estimates. The texture feature subset selected to estimate each forest variable and the accuracy of the derived predictions were inherent to the studied site and the used image (sensor, date). A model that was determined from a given field-measured dataset and a given image gave poor estimation results when applied to another site or another image. This can be explained by various uncontrolled parameters such as seasonal variations (i.e. phenology stages, solar elevation range) or the viewing angle during image acquisition, which affects the interactions of

radiation with a given forest canopy as measured by the sensor, and thus image texture, as previously mentioned.

The differences in performance between the images acquired on Tagon in June and August were quite small. However, this study did not consider images acquired on the same site with very contrasting various conditions (winter to summer, nadir to very oblique viewing). Further work should be carried out to identify the best image conditions to ensure a satisfactory prediction performance, i.e. a texture that is strongly related to the tree canopy structure.

#### 4.2.3. Assessment of prediction performance

This last section focuses on the validation of our method for modeling a given forest variable using a multi-variable solution. This assessment was made on Tagon site where various field observations are available ( $n = 111$ ) and are assumed to be representative of the forest structure variability in maritime pine stands over the study area.

The field samples were split into two independent datasets: a training set to find the best regression model using Random PRIF method on the MS-Pan image data and a test set to calculate the accuracy (i.e. RMSE) of the forest variable predicted from this model. The prediction robustness was assessed by repeating the modeling step on different training sets of increasing size. Thus the initial field samples were split into two subsets of same distribution (the samples were sorted with respect to their crown diameter and were assigned alternatively either as training or as test sample). The first subset contained 71 samples from which all training sets were randomly sampled with replacement. The same sampling was drawn for both dates. The second subset which contained the remaining 40 samples was the only one test set. The RMSE was computed from the predictions calculated on test set while increasing the size of the training set from 15 to 55 samples. The RMSE was averaged over 30 runs.

Fig. 7 illustrates the results obtained on June and August images for Cd (crown diameter). It shows that the average RMSE decreases significantly while increasing the training set size. Average RMSE decreases by approximately 40% from 15 observations within the training set to 45 observations (i.e. from 2.5 m to 1.5 m and from 3.2 m to 1.4 m for June and August images, respectively). Then it reaches a plateau with a value close to the LOO RMSE previously calculated ( $\sim 1.1$  m for both dates). On the plateau, the maximal standard deviation is approximately 0.50 m. These results show the minimum number of samples required to take into account the variability of the forest structure that is common in the Landes de Gascogne forest and to ensure a good predicting performance of main forest variables. For Tagon site, using an equal number of 40 samples in both training and test sets is a good trade-off to guarantee for both dates a model estimating Cd with a maximum error of  $\sim 2$  m. With such a model, it would be possible to identify several classes of crown diameters ranging from 1 m or less to about 11 m. Nevertheless, further work should be undertaken with more in-field samples in order to refine the estimates of the expected maximal errors on the different forest variables.

## 5. Conclusion

This work provided an automated method to retrieve forest structure variables using texture information extracted from VHR optical satellite images. An automated feature selection process based on the use of image test frames of forest samples was investigated. This method allows a fast assessment of thousands of descriptors, exploring a wide range of parameter values, to optimize spatial texture parameter settings. The best image texture features are selected inherently leading to multi-scale texture

descriptors via statistical modeling. Five typical forest structure variables were successfully modeled. Our forest variable retrieval undergoes both multi-resolution and multi-scale texture analysis, combining panchromatic and/or multispectral features with different spatial parameter settings. The multicollinearity problem, encountered in multi-linear regression, is addressed carefully using the VIF criterion. We propose a new variable subset selection scheme, we called Random PRiF, based on a global optimization via random sampling in feature space. This approach involves a new statistical criterion we named PRiF, combining both VIF and PRESS criteria. Comparing our method to two well-established feature subset selection methods (LARS stepwise and PCA), it turned out the most convincing, significantly improving the quality of estimation of all forest structure variables for both sites. As future work, other feature subset selection methods such as tree-based regression for instance should be tested.

Exploring texture feature combinations from both panchromatic and multispectral datasets leads to the best prediction performances. This approach could be extended to other applications involving VHR image texture. The whole protocol we have introduced can be easily applied to any other forest using site-specific test frames. Finally, the use of Pléiades images showed the great potential of this sensor for forest applications.

Future work should be conducted on various acquisition conditions on the same site (winter to summer, nadir to very oblique viewing) to identify the best image conditions to ensure a satisfactory prediction performance.

## Acknowledgements

This research was funded with grants from Conseil Régional d'Aquitaine and CNES (Centre National d'Etudes Spatiales). The VHR images were provided by the Orfeo program set up by CNES in order to prepare the use of Pléiades images. The authors are very grateful to Bernard Issenhut (INRA, Unité expérimentale Forêt Pierron) for the measurements in the field and the good quality of the provided forest data. Further thanks to Christian Germain from IMS Lab (Bordeaux), Jordi Inglada from CESBIO (Toulouse) and Jean-Pierre Wigneron from INRA (Bordeaux) for discussions.

## References

- Allen, D.M., 1974. The relationship between variable selection and prediction. *Technometrics* 16 (125), 125–127.
- Barbier, N., Couteron, P., Proisy, C., Malhi, Y., Gastellu-Etchegorry, J.P., 2010. The variation of apparent crown size and canopy heterogeneity across lowland Amazonian forests. *Global Ecol. Biogeogr.* 19, 72–84.
- Beguet, B., Chehata, N., Boukir, S., Guyon, D., 2012. Retrieving forest structure variables from Very High Resolution satellite images using an automatic method. *ISPRS Ann. Photogramm., Remote Sens. Spatial Inform. Sci.* 1-7, 1–6.
- Beguet, B., Boukir, S., Guyon, D., Chehata, N., 2013. Modelling-based feature selection for classification of forest structure using very high resolution multispectral imagery, SMC'2013, IEEE Int. Conf. on Systems, Man, and Cybernetics, Manchester, UK, pp. 4294–4299.
- Boukir, S., Orny, C., Chehata, N., Guyon, D., Wigneron, J., 2013. Détection de changements structurels sur des images satellite haute résolution. Application en milieu forestier., *Traitement du signal* 6, 401–429.
- Castillo, M.A., Ricker, M., Jong, B.H.J.D., 2010. Estimation of tropical forest structure from Spot-5 satellite images. *Int. J. Remote Sens.* 31 (10), 2767–2782.
- Chehata, N., Orny, C., Boukir, S., Guyon, D., 2011. Object-based forest change detection using high resolution satellite images, In: PIA'2011, ISPRS Photogrammetric Image Analysis Conference, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Munich, Germany, vol. 38(3), pp. 49–54.
- Chehata, N., Orny, C., Boukir, S., Guyon, D., 2014. Object-based change detection in wind-storm damaged forest using high resolution multispectral images. *Int. J. Remote Sens.* 35, 4758–4777.
- Coburn, C.A., Roberts, A.C.B., 2004. A multiscale texture analysis procedure for improved forest stand classification. *Int. J. Remote Sens.* 25, 4287–4308.
- Conners, R., Trivedi, M., Harlow, C., 1984. Segmentation of a high-resolution urban scene using texture operators. *Comput. Vision Graphics Image Process.* 25 (3), 273–310, ISSN 0734189X, doi:10.1016/0734-189X(84)90197-x.
- Couteron, P., Pelissier, R., Nicolini, E.A., Paget, D., 2005. Predicting tropical forest stand structure parameters from Fourier transform of very high-resolution remotely sensed canopy images. *J. Appl. Ecol.* 42, 1121–1128.
- Culbert, P.D. et al., 2009. The impact of phenological variation on texture measures of remotely sensed imagery. *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens* 2 (4), 299–309.
- Dell'Acqua, F., Gamba, P., 2003. Texture-based characterization of urban environments on satellite SAR images. *IEEE Trans. Geosci. Remote Sens.* 41 (1), 153–159.
- Efron, B., Hastie, T., Tibshirani, I.J.R., 2004. Least angle regression. *Ann. Statist.* 32 (2), 407–499.
- Feng, Y., Li, Z., Tokola, T., 2010. Estimation of stand mean crown diameter from high-spatial-resolution imagery based on a geostatistical method. *Int. J. Remote Sens.* 31 (2), 363–378.
- Franklin, S.E., Maudie, A.J., Lavigne, M.B., 2001. Using spatial co-occurrence texture to increase forest structure and species composition classification accuracy. *Photogramm. Eng. Remote Sens.* 67 (7), 849–855.
- Gaetano, R., Scarpa, G., Poggi, G., 2009. Hierarchical texture-based segmentation of multiresolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 47, 2129–2141.
- Gomez, C., Wulder, M.A., Montes, F., Delgado, J.A., 2012. Modeling forest structural parameters in the mediterranean pines of central Spain using Quickbird-2 imagery and classification and regression tree analysis (CART). *Remote Sens.* 4, 135–139.
- Gonzales, R.C., Woods, R.E., 2008. *Digital Image Processing*. Pearson Education Inc. Gromping, U., 2006. Relative importance for linear regression in R: the package relaimpo. *J. Statist. Software* 17, 1–27.
- Guyon, D., Riou, J., 1996. Estimation de caractéristiques forestières à partir d'images à haute résolution spatiale (Spot5). *BUL.S.F.P.T.* 141, 46–50.
- Haan, C.T., 2002. *Statistical methods in Hydrology*, second ed. Iowa State University Press, Ames, Iowa.
- Haralick, R.M., Shanmugan, K., Dinstein, I., 1973. Texture features for image classification. *IEEE Trans. Syst. Man Cybern.*, 610–621.
- Hyyppä, J., Hyyppä, H., Inkinen, M., Engdahl, M., Linko, S., Zhu, Y.H., 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stands attributes. *Forest Ecol. Manage.* 128 (1–2), 109–120.
- Kayitakire, F., Hamel, C., Defourny, P., 2006. Retrieving forest structure variables based on image texture analysis and IKONOS-2 imagery. *Remote Sens. Environ.* 102, 390–401.
- Kutner, M.H., Nachtsheim, C.J., Neter, J., 2004. *Applied Linear Regression Models*, fourth ed. McGraw-Hill Irwin.
- Maillard, P., 2003. Comparing texture analysis methods through classification. *Photogramm. Eng. Remote Sens.* 69 (4), 357–367.
- Manly, B.F.J., 1994. *Multivariate Statistical Methods: A Primer*. Chapman & Hall, London, UK.
- Murray, H., Lucieer, A., Williams, R., 2010. Texture-based classification of sub-Antarctic vegetation communities on Heard Island. *Int. J. Appl. Earth Observation Geoinformation* 12, 138–149.
- Ozdemir, I., Karnieli, A., 2011. Predicting forest structural parameters using the image texture derived from WorldView-2 multispectral imagery in a dryland forest, Israel. *Int. J. Appl. Earth Observation Geoinformation* 13, 701–710.
- Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sens. Environ.* 113 (6), 1276–1292.
- Porté, A., Bosc, A., Champion, I., Loustau, D., 2000. Estimating the foliage area of Maritime pine (*Pinus pinaster* Ait.) branches and crowns with application to modelling the foliage area distribution crown. *Ann. For. Sci.* 57, 73–86.
- Proisy, C., Couteron, P., Fromard, F., 2007. Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images. *Remote Sens. Environ.* 109, 379–392.
- Regniers, O., DaCosta, J.-P., Grenier, G., Germain, C., Bombrun, L., 2013. Texture based image retrieval and classification of very high resolution maritime pine forest images. In: *Proceedings IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.
- Ruiz, L.A., Fdez-Sarria, A., Recio, J.A., 2004. Texture feature extraction for classification of remote-sensing data using wavelet decomposition: a comparative study. In: *20th ISPRS Congress*.
- Soe, W.M., Tyler, S.N.L.J.M., 2004. Wavelets for urban spatial feature discrimination: comparisons with fractal, spatial autocorrelation, and spatial co-occurrence approaches. *Photogramm. Eng. Remote Sens.* 70 (7), 803–812.
- Song, C., Dickinson, M.B., Su, L., Zhang, S., Yaussey, D., 2010. Estimating average tree crown size using spatial information from Ikonos and QuickBird images: across-sensor and across-site comparisons. *Remote Sens. Environ.* 114, 1099–1107.
- St-Onge, B.A., Cavayas, F., 1997. Automated forest structure mapping from high resolution imagery based on directional semivariogram estimates. *Remote Sens. Environ.* 61, 82–95.
- Trias-Sanz, R., Stamon, G., Louchet, J., 2008. Using colour, texture, and hierarchical segmentation for high-resolution remote-sensing. *ISPRS J. Photogramm. Remote Sens.* 63, 156–168.
- Tuominen, S., Pekkarinen, A., 2005. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. *Remote Sens. Environ.* 94, 256–268.

- Tuttle, E.M., Jensen, R.R., Formica, V.A., Gonser, R.A., 2006. Using remote sensing image texture to study habitat use patterns: a case study using the polymorphic white-throated sparrow (*Zonotrichia albicollis*). *Global Ecol. Biogeogr.* 15, 557–559.
- Van Coillie, F.M., Verbeke, L.P., Wulf, R.R.D., 2007. Feature selection by genetic algorithms in object-based classification of (IKONOS) imagery for forest mapping in Flanders, Belgium. *Remote Sens. Environ.* 110 (4), 476–487, ISSN 0034-4257.
- Wolter, T., Townsend, P.A., Sturtevant, B.R., 2009. Estimation of forest structural parameters using 5 and 10 meter Spot-5 satellite data. *Remote Sens. Environ.* 113, 2019–2036.
- Wood, E.M., Pidgeon, A.M., Radeloff, V.C., Keuler, N.S., 2012. Image texture as a remotely sensed measure of vegetation structure. *Remote Sens. Environ.* 121, 516–526.
- Wulder, M.A., Drew, E.F.L., Franklin, S.E., Lavigne, M.B., 1998. Aerial image texture information in the estimation of Northern deciduous and mixed wood forest leaf area index (LAI). *Remote Sens. Environ.* 64, 64–76.
- Wunderle, A.L., Franklin, S.E., Guo, X.G., 2007. Regenerating boreal forest structure estimation using Spot-5 pansharpened imagery. *Int. J. Remote Sens.* 28 (19), 4351–4364.

## Annexe F

**Classification of forest structure  
using very high resolution  
Pléiades image texture.**

# CLASSIFICATION OF FOREST STRUCTURE USING VERY HIGH RESOLUTION PLEIADES IMAGE TEXTURE

Beguet B.<sup>1,2</sup>, Chehata N.<sup>1,3</sup>, Boukir S.<sup>1</sup>, Guyon D.<sup>2</sup>

<sup>1</sup> G&E Laboratory (EA 4592), IPB / University of Bordeaux,  
1 Allée F. Daguin, 33607 Pessac Cedex, France.

<sup>2</sup> INRA, UMR1391 ISPA, 33140 Villenave d'Ornon, France.

<sup>3</sup> LISAH UMR 144, IRD, El Menzah 4, Tunis, Tunisia.

## ABSTRACT

The potential of very high resolution Pléiades image texture for forest structure mapping was assessed on maritime pine stands in south-western France. A preliminary step showed that multi-linear regressions allow a reliable prediction of forest variables (such as crown diameter or tree height) from a set of features automatically selected among a huge number of Haralick texture features with various spatial parameterizations. In a second step, to assess Pléiades image texture contribution for classification, Random Forests (RF) classification was performed to discriminate four forest structure classes from recent reforestation to mature stand. Two texture feature selection strategies are compared: (1) the previous regression-based modelling using in situ tree measurements (2) the RF-variable importance using a visual photo-interpretation. Both methods produced comparable classification accuracies. The results highlight the contribution of processes automation and the need for using both Pléiades image resolutions (panchromatic and multispectral) to derive the best performing texture features.

**Index Terms**—Forest, Texture, Feature Selection, Classification, Pléiades

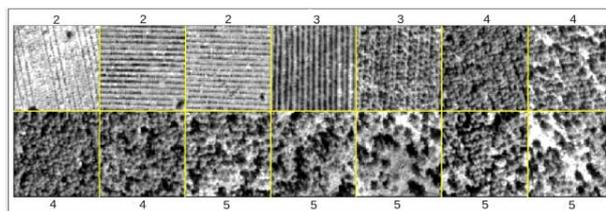
## 1. INTRODUCTION

Very High resolution (VHR) images allows a detailed description of forest structure (tree distribution and size) at the stand level, by exploiting the strong relationship between tree spatial structure and image texture when the pixel size is smaller than the tree dimension. This expectation meets the strong need for spatial forest resources inventory at the stand level and its changes due to forest management, land use or catastrophic events. Our objective is to assess the potential of Pléiades images to map the forest structure. The approach was conducted in two complementary steps:

- The first concerns the estimation of the main forest structure variables (crown diameter, stem diameter, height, density or tree spacing) from the image texture. On these bases, different forest structure classes could be considered and a pixel-based image classification could be processed in order to produce the finest possible spatial information on forest structure.
- The second aims to classify these forest structure classes when no ground measurements are available. This issue is explored by proposing an alternative methodology.

## 2. MATERIAL

The study covers a 80-km<sup>2</sup> site located in south-western France (Tagon–Marcheprime, between Bordeaux and the Bay of Arcachon) within the largest European maritime pine (*Pinus pinaster* Ait.) forest. It was conducted using two Pléiades image sets in panchromatic (Pan, resampled spatial resolution= 0.50m) and multispectral bands in blue, green, red and near infrared (MS, 2m-), acquired on 26 June and 8 August 2012.



**Figure 1:** Forest image samples corresponding to ground measurements (ordered by structure class). Panchromatic Pléiades image, 26 June 2012. Class 2 = crown diameter < 1m, tree rows are visible ; Class 3 = crown diameter is between 1m and 2.5m, tree rows are visible ; Class 4 = crown diameter is between 2.5m and 4.5m, tree rows are no more visible; Class 5 = crown diameter > 4.5m.

Some forest structure variables were measured during summer 2012 on a large number of maritime pine stands (n=111) at the end of the tree pine growth time: tree height (varying from 1.7m to 26.2m), crown diameter (0.77m - 10.69m), diameter at breast height (0.02m - 0.56m), density (150tree/ha - 6729tree/ha) and tree spacing (1.31m - 8.77m). The trees sampled in each stand were included in a 20m × 20m square plot representative of the forest structure in the 80m × 80m area that encloses it.

For each sampled stand we considered an image sample covering a square area centered on a field measurement plot with a slightly larger width (60 meters, i.e. 120 pixels in Pan and 30 pixels in MS). Some image samples are given in figure 1 as a non-exhaustive illustration of the variability of forest structure and image texture.

### 3. METHODS

We used texture features derived from the co-occurrence matrices (GLCM): the eight most common Haralick's texture features [1], and also mean and variance (first order statistics). They are commonly used for forest studies [2,3,4]. An automatic modeling procedure based on multiple linear regression of each forest variable by the texture features [3] is proposed and evaluated based on image samples corresponding to the in-situ measurements. The approach consists in selecting as predictors the most efficient combination of texture features calculated in Pan (and then resampled to MS resolution) and MS bands: best R<sup>2</sup> and RMSE while ensuring a minimal collinearity and a balanced contribution importance [3]. This allowed us to

explore a wide range of parameters of GLCM (window size, displacement and orientation) and thus to optimize them automatically.

Five forest structure classes, representative of crown dimension and spatial distribution are defined. The first class, corresponding to clear cuts and stands with trees younger than 3 years is not treated in this study (as no measurements were done on these stands, where the pine trees are less tall than non-forest species). Classes 2 to 5 are defined in figure 1.

For pixel-level classification, two feature selection methods were compared: the previous modelling-based approach using 111 stands and a Random Forest (RF) ranking using variable importance [5,6]. The latter, which could be proposed when no ground measurements are available, is based on 40 plots among 111 and their forest structure classes are recognized by visual photo-interpretation of the Pan image. Then, all texture features with various parametrizations are put into the RF classifier, the best parameterizations are selected by RF variable importance [4].

Finally, the classification is processed by Random Forests [7,8] using 100 trees for the two sets of selected features. Two independent sets of image samples are used for learning (n=80) and for test (n=80). They are also independent of the samples used for feature selection and are selected by photointerpretation of the Pan image. The classification performances of both feature selection methods are compared.

		Classification						
		2	3	4	5	Rowtotal	Commission Error %	
Reference	2	14416	3174	233	177	18000	24,86	
	3	1392	13115	3395	97	18000	27,14	
	4	791	864	14821	1524	18000	17,66	
	5	5	1014	2538	14443	18000	19,76	
	Column total	16604	18167	20988	16241	72000		
Omission Error %		13,18	27,81	28,38	11,07			
O.A =		78,88						
Kappa =		71,84						

**Table 1:** Classification confusion matrix obtained with the texture features selected by RF variable importance.

		Classification						
		2	3	4	5	Rowtotal	Commission Error %	
Reference	2	13899	3587	379	135	18000	22,78	
	3	2132	12495	2716	657	18000	30,58	
	4	900	1200	12201	3699	18000	32,21	
	5	0	444	2191	15365	18000	14,63	
	Column total	16931	17726	17487	19856	72000		
Omission Error %		17,9	29,51	30,22	22,61			
O.A =		74,94						
Kappa =		66,59						

**Table 2:** Classification confusion matrix obtained with the texture features selected by the regression-based modelling of forest variables.

#### 4. RESULTS AND DISCUSSION

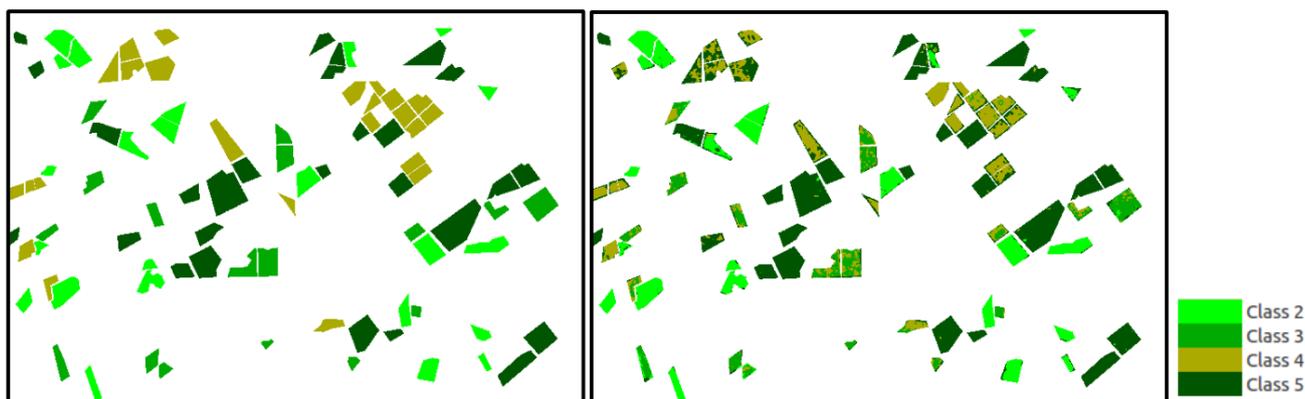
The modelling results show that the best performing regression models combine panchromatic and multispectral (different spatial and spectral resolutions) texture features. The prediction accuracies (RMSE) obtained using the five best performing texture features as predictors of forest variables are similar for both dates:  $\sim 1.1\text{m}$  for crown diameter,  $\sim 3\text{m}$  for tree height,  $\sim 0.9\text{m}$  for tree spacing,  $\sim 586\text{tree/ha}$  for density and  $\sim 0.066\text{m}$  for diameter at breast height. These results are very satisfactory. Therefore we can expect to discriminate several forest structure classes from image classification.

Both feature selection methods used for classification provide a subset of 50 texture features to which the four original spectral bands are added. The classification results are presented in Table 1 and Table 2. Only results from the June image are presented (they are similar to August image). The overall classification accuracy reaches 78.8% using the RF importance feature selection, improving the results by almost 4% with comparison to the modelling based feature selection. However, due to the complexity of forest structure

and photo-interpretation labelling, these results are considered as similar. This indicates the potential of our methodology to identify forest structure classes when no ground measurements are available.

We notice that the larger error rates (both in commission and omission) occur in the intermediate classes (3 and 4) for both feature selection methods. These confusions can be observed in Figure 2. They highlight the complexity of forest structure classification, as the class borders are fuzzy since forest variable distributions are continuous. Figure 2 also shows that most of the classification errors appear on stand borders which is inherent to a spatial descriptor such as texture. A regularization step should enhance the classification quality and it will be investigated in order to produce an operational forest structure map.

The introduction of the class 1 (clear cuts and stands with trees younger than 3 years), which seems very different from others both in radiometry and texture, should increase the classification accuracy, and it will be done in further work.



**Figure 2:** Reference map (left) and classification map (right) obtained with the texture features selected by RF ranking of forest variables on a part of the study area.

#### 5. CONCLUSION

In conclusion, the results show the potential of Pléiades image texture to estimate the main forest structure variables and to map forest structure.

The proposed methodologies allow us to optimize the use of VHR Pléiades images, calculating texture features with various spatial parameterizations on both types of spatial and spectral resolution (50cm in Panchromatic, 2m in multispectral mode) before finding their best combinations

for both forest variable estimation and forest structure classification.

The texture features representative of forest variables allow us to classify forest structure with a satisfying accuracy into four classes from recent reforestation ( $>3$  years) to mature stand. When no ground measurements are available, the proposed methodology appears to be a satisfying alternative to drive a forest structure classification task.

The robustness of the proposed methodologies will be assessed using a third Pléiades image, acquired in winter (20 February 2013, i.e. with a lower sun and a dry understory

vegetation). In addition, their application to multi-annual images would allow us to assess their ability to detect and map strong forest changes such as forest cuts, urban sprawl or storm damage.

## 6. ACKNOWLEDGEMENTS

This research was funded with grants from Conseil Régional d'Aquitaine and CNES (Centre National d'Etudes Spatiales). The Pléiades images were provided by the Orfeo thematic program set up by CNES. The work is based on the use of OrfeoToolBox [9]. The authors are very grateful to Bernard Issenhut (INRA, Unité expérimentale Forêt Pierroton) for the measurements in the field and the good quality of the provided forest data. Further thanks to Christian Germain from MS Lab (Bordeaux), Jordi Inglada from CESBIO (Toulouse) and Jean-Pierre Wigneron from INRA (Bordeaux) for discussions.

## 7. REFERENCES

- [1] R.M. Haralick, K. Shanmugan and I. Dinstein, "Texture features for image classification", IEEE Transactions Systems, Man and Cybernetics, pp. 601-621, 1973.
- [2] F. Kayitakire, C. Hamel and P. Defourny, "Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery", Remote Sensing of Environment 102, pp. 390-401, 2006.
- [3] B. Beguet, N. Chehata, S. Boukir and D. Guyon, "Retrieving forest structure variables from very high resolution satellite images using an automatic method", ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences I-7, pp. 1-6, 2012.
- [4] M. A. Castillo, M. Ricker, and B. H. J. De Jong, Estimation of tropical forest structure from spot-5 satellite images. International Journal of Remote Sensing, vol. 31-10, pp. 2767-2782, 2010.
- [5] R. Genuer, J.M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests", Pattern Recognition Letters vol. 31, pp. 2225-2236, 2010.
- [6] Dash M., Liu H., Feature Selection for Classification, Intelligent Data Analysis, vol. 1, p. 131-156, 1997.
- [7] L. Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- [8] M. Pal, Random forest classifier for remote sensing classification. International Journal of Remote Sensing, vol. 26(1), pp. 217-222, 2005.
- [9] J. Inglada and E. Christophe, The Orfeo toolbox remote sensing image processing software. IEEE International Geoscience and Remote Sensing Symposium, vol. 4, pp. IV-733-IV-736, 2009.



## Annexe G

Quantification et cartographie de la structure forestière à partir de la texture des images Pléiades.

# QUANTIFICATION ET CARTOGRAPHIE DE LA STRUCTURE FORESTIERE A PARTIR DE LA TEXTURE DES IMAGES PLEIADES

Benoît Beguet<sup>1,2</sup>, Nesrine Chehata<sup>1</sup>, Samia Boukir<sup>1</sup>, Dominique Guyon<sup>2</sup>

1 : ENSEGD/Université de Bordeaux, laboratoire G&E (EA 4592), Pessac, benoit.beguet@ipb.fr

2 : INRA, UMR1391 ISPA, Villenave d'Ornon

## Résumé

Cette étude montre le potentiel de l'information texturale des images à très haute résolution spatiale Pléiades pour la quantification et la cartographie de la structure forestière des peuplements de pin maritime du sud-ouest de la France (massif forestier landais). Une première étape montre qu'il est possible d'estimer, par régressions linéaires multiples, les variables de structure forestière (comme le diamètre des couronnes ou la hauteur des arbres) à partir d'un ensemble d'attributs de texture automatiquement sélectionnés parmi un grand nombre de paramétrages possibles. La classification de l'image est ensuite effectuée en utilisant l'algorithme des forêts aléatoires (RF) pour discriminer cinq classes de structure forestière avec une approche hiérarchique. L'importance de variable des RF est utilisée pour la sélection des attributs de texture. Les résultats montrent l'intérêt de l'automatisation du processus et de l'utilisation conjointe des deux résolutions des images Pléiades (mode panchromatique et mode multispectral) pour dériver les attributs de texture les plus performants pour détecter de fines variations de structure forestière.

**Mots-clés :** Classification, Sélection de variables, forêts aléatoires, Texture, Forêt, Pléiades, Très haute résolution spatiale

## Abstract

*The potential of very high spatial resolution Pléiades image texture for forest structure quantification and mapping was assessed on maritime pine stands in south-western France. A preliminary step showed that multi-linear regressions allowed a reliable prediction of forest variables (such as crown diameter or tree height) from a set of features automatically selected among a huge number of texture features with various spatial parameterizations. The image classification was performed using the Random Forests (RF) ensemble classifier in order to discriminate five forest structure classes using a hierarchical approach. The RF-variable importance is used for texture feature selection. The results highlight the contribution of process automation and the need for a joint use of both Pléiades image resolutions (panchromatic and multispectral) to derive the best performing texture features.*

**Keywords:** Classification, Feature selection, Random forest, Texture, Forestry, Pléiades, Very high spatial resolution

## 1. Introduction

Les images à Très Haute Résolution (THR) spatiale donnent accès à une description de la structure forestière, au sens de la distribution des dimensions des arbres, en exploitant sa forte relation avec la texture des images quand la taille du pixel est inférieure à la dimension des arbres. Cette perspective répond au fort besoin d'inventaires fins et spatialisés de la ressource forestière et de ses changements dus à la gestion forestière, à une modification dans l'usage du sol ou suite à des événements catastrophiques (type tempête).

Notre étude a pour but d'évaluer le potentiel des images THR Pléiades et tout particulièrement de leur information texturale pour caractériser et cartographier de manière fine la structure forestière. Sont ciblés les peuplements forestiers traités en futaie régulière pure, c'est à dire constitués d'arbres d'une seule espèce et du même âge ou presque (peuplement équienne), à l'instar de ceux de pin maritime du massif forestier landais (sud-ouest France) qui, couvrant près d'un million d'hectares, ont une forte importance économique. Nous avons deux objectifs complémentaires. Le premier vise l'estimation de variables de structure forestière (grandeurs dendrométriques telles que diamètre des couronnes, hauteur, diamètre des troncs, densité ou espacement

des arbres) via une modélisation de leur relation avec des attributs de texture. Si les erreurs d'estimation ne sont pas trop fortes devant l'étendue de ces variables, on peut espérer identifier plusieurs classes de structure forestière couvrant tout le cycle de développement des peuplements. Le second objectif vise alors à identifier différents stades de développement des peuplements (du renouvellement à la maturité) à l'aide d'une classification d'image au niveau pixel de façon à produire une information forestière spatialisée la plus fine possible.

L'analyse de texture d'image THR appliquée aux problématiques forestières peut être divisée en trois familles principales (Maillard, 2003) : les méthodes basées sur les ondelettes (Van Coillie *et al.*, 2007 ; Regniers *et al.*, 2013) ou sur la transformée de Fourier (Couteron *et al.*, 2005 ; Proisy *et al.*, 2007), l'utilisation des variogrammes (St-Onge *et al.*, 1997 ; Song *et al.*, 2010) et enfin l'extraction des attributs de texture issus des matrices de co-occurrence (GLCM).

Nous avons choisi d'exploiter la texture sur la base des matrices de co-occurrence qui sont très utilisées (Kayitakire *et al.*, 2006 ; Castillo *et al.*, 2010 ; Boukir *et al.*, 2013 ; Chehata *et al.*, 2014 ; Beguet *et al.*, 2014). Leur principal inconvénient est le grand nombre de valeurs possibles des paramètres spatiaux à explorer. Pour remédier à ce problème, nous proposons une approche automatique qui consiste à sélectionner parmi

un très grand nombre d'attributs de texture (avec une grande diversité de paramétrage), un ensemble réduit d'attributs qui sont complémentaires entre eux et adaptés à l'application considérée. Ceci permet d'exploiter au maximum le potentiel des images Pléiades et de combiner les deux résolutions (panchromatique et multispectrale). Une méthodologie de sélection d'attributs de texture propre à chacun des deux objectifs visés est développée.

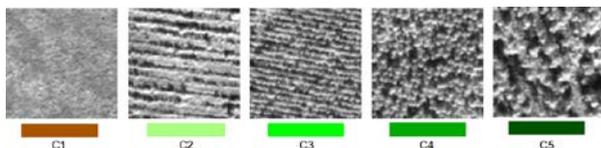
L'étude a été réalisée sur des peuplements de pin maritime (*Pinus pinaster* Ait.) du massif forestier landais. Pour tester la robustesse des méthodes développées et évaluer l'apport de l'imagerie Pléiades, elle s'appuie sur deux jeux d'images acquis sur le même site à deux dates différentes.

## 2. Site d'étude et données

Le site d'étude (Tagon-Marcheprime) couvre une surface de 80 kmentre Bordeaux et Arcachon. Comme généralement dans le massif landais, la sylviculture du pin maritime y est dynamique : les peuplements sont purs et équiennes ; ils sont renouvelés par semis en ligne ou plantation ; vers l'âge de 50 ans ils sont exploités en coupe rase ; entre temps ils sont éclaircis périodiquement (Maugé, 1987).

Deux jeux d'images Pléiades ont été acquis les 26 juin et 8 août 2012 en modes panchromatique (Pan, produit image ré-échantillonné à 0,50 m) et multispectral (MS, ré-échantillonné à 2 m, quatre bandes spectrales : bleu, vert, rouge et proche infra-rouge).

Une campagne de mesures *in situ* a été menée durant l'été 2012 à la fin de la période annuelle de croissance sur un grand nombre de peuplements de pin maritime (n=111) couvrant une large gamme de structure sur quasiment tout le cycle de développement (de l'âge de 3-4 ans à la maturité). Les variables mesurées sont : la hauteur des arbres (Ht, variant de 1,7 à 26,2 m), le diamètre des couronnes (Cd, de 0,77 à 10,69 m), le diamètre des troncs à 1,30m (Dbh, de 0,02 à 0,56 m), la densité des tiges (Nah, de 150 à 6729 arbres/ha). Un indicateur de l'espacement entre les arbres est calculé par transformation non linéaire de la densité (Sp, de 1,31 à 8,77 m). Les arbres mesurés dans chaque peuplement sont inclus dans un carré de 20 m x 20 m représentatif de la structure forestière dans un voisinage englobant de 80 m x 80 m. Pour chaque peuplement, un échantillon d'image centré sur la placette d'une taille de 60 m x 60 m est considéré (soit 120 pixels pour Pan et 30 pixels pour MS).



**Figure 1 :** Imagettes panchromatiques et classes de structure forestière associées.

C1 : pas d'arbres visibles, en mutation (coupe rase, reboisement en cours ou récent, âge  $\leq 3$  ans environ) ;  
 C2 : arbres visibles, rangs visibles, Cd < 1 m ;  
 C3 : 1 m < Cd < 2,5 m, rangs visibles  
 C4 : 2,5 m < Cd < 4,5 m, rangs non visibles ;  
 C5 : Cd > 4,5 m.

En complément, 200 autres peuplements ont été reconnus par photo-interprétation visuelle des images en panchromatique pour constituer une base de données de référence pour la classification. Cinq classes de structure forestière ont été considérées pour décrire les stades de développement, elles sont représentatives de la dimension des couronnes (Cd) et de la distribution spatiale des arbres. Elles sont décrites en Figure 1.

## 3. Méthodologie

Les attributs de texture utilisés sont deux du premier ordre, moyenne et variance locale, et huit du second ordre, dérivés des matrices de co-occurrence (GLCM) et définis dans Haralick *et al.* (1973). Ces derniers (énergie, entropie, corrélation, inverse différence, inertia, cluster shade, cluster prominence, Haralick's Correlation) sont les plus couramment utilisés en télédétection forestière (Kayitakire *et al.*, 2006 ; Castillo *et al.*, 2010 ; Beguet *et al.*, 2014). Les attributs ont été calculés avec une vaste gamme de paramétrage spatial : rayon de la fenêtre de calcul ( $r$ ), orientation ( $o$ ), magnitude du déplacement ( $d$ ). Le nombre de niveaux de quantification a été fixé à 8 pour tous les attributs de texture. Ces attributs sont calculés sur chacune des bandes spectrales MS ainsi que sur la bande panchromatique, avec des gammes de paramètres adaptés présentées dans le Tableau 1. Les attributs Pan sont ré-échantillonnés après calcul à la résolution MS.

	$r$ (pixel)	$d$ (pixel)	$o$ (degrés)
<b>Pan</b>	5-25 ; pas : 5	1-10 ; pas : 1	0-135 ; pas : 45
<b>MS</b>	3-12 ; pas : 3	1-4 ; pas : 1	0-135 ; pas : 45

**Tableau 1 :** Gammes de variation des paramètres spatiaux des GLCM.

L'idée principale, commune aux deux objectifs, est d'optimiser de manière automatique le paramétrage des attributs de texture (Beguet *et al.*, 2014). Cela est fait dans les étapes de sélection de variables parmi tous les d'attributs de texture possibles, plusieurs milliers, correspondant aux dix attributs définis précédemment, calculés sur toutes les données Pan et MS (soit deux résolutions spatiales et 5 bandes spectrales) pour chaque combinaison de paramètres spatiaux.

### 3.1 Modélisation des variables forestières

Des régressions linéaires multiples sont appliquées entre chacune des variables forestières d'intérêt et un ensemble d'attributs de texture image. Une méthode de sélection de variable adaptée aux régressions multiples est donc appliquée afin de trouver un sous-ensemble d'attributs de texture le plus représentatif de la variable de structure forestière concernée (en termes de R et de RMSE, erreur quadratique moyenne, obtenue par validation croisée, c'est-à-dire prédiction d'une observation par modélisation sur les 110 autres (notée LOO RMSE)). Le sous-ensemble d'attributs obtenu se doit d'être à la fois équilibré, en terme d'importance relative, et exempt de colinéarité (problème très perturbant en régression linéaire multiple). La méthode de sélection de variables nommée « Random PRiF », mise au point dans une précédente étude (Beguet *et al.*, 2014), a été appliquée. Cette méthode permet de sélectionner un ensemble d'attributs complémentaires en minimisant grandement la colinéarité. Pour chaque

variable de structure forestière, un ensemble de 5 attributs de texture est sélectionné parmi l'ensemble des attributs de texture (MS et Pan) pour modéliser les différentes variables forestières.

### 3.2 Classification des stades de développement

La classification a été conduite de manière hiérarchique. Il s'agit, dans un premier temps, de discriminer la classe C1 (pas d'arbres visibles), qui correspond aux stades de mutation de la forêt avec un taux de couvert forestier quasi nul, des autres classes C2 à C5 qui correspondent aux stades ultérieurs de développement des peuplements (âge > 3 ans). Cette première classe C1 est donc très différente des autres classes. Cependant elle présente une très forte variabilité dans la composition et le développement de la végétation (et donc en radiométrie et en texture), ce qui rend sa reconnaissance difficile lorsque l'on cherche à différencier les cinq classes de manière directe. Une fois cette première discrimination effectuée, la classification des stades C2 à C5 est effectuée. L'approche hiérarchique permet d'identifier les surfaces C1 dans un premier temps, celles-ci sont ensuite masquées pour calculer à nouveau la texture uniquement sur les autres surfaces forestières. La dynamique radiométrique à traiter est alors réduite (concentrée sur les pixels de forêt), ce qui conduit à la création d'attributs GLCM plus adaptés, cela pour un même nombre de niveaux de quantification. Les deux résultats sont ensuite regroupés pour former la carte finale des 5 classes.

Deux étapes de sélection de variables sont alors nécessaires, une première étape pour trouver les attributs de texture qui discriminent le plus C1 des autres classes, puis une seconde étape de sélection de variables pour déterminer un ensemble d'attributs performant pour la discrimination entre les autres stades. La méthode de sélection de variable utilisée consiste à ordonner les attributs de texture en fonction de l'importance de variable mesurée dans un processus de classification par forêts aléatoires (ou Random Forests, RF) (Breiman, 2001). L'importance d'une variable est mesurée par la perte en précision de classification lorsqu'elle est modifiée aléatoirement (simulée comme absente). Cette sélection est faite sur un ensemble de 40 échantillons (choisis parmi les 111 peuplements mesurés *in situ*), ceci pour les deux étapes, à partir de l'ensemble complet des attributs de texture (MS et Pan).

La classification de l'image est faite par l'algorithme des forêts aléatoires (Breiman, 2001) en utilisant 100 arbres. Deux jeux d'échantillons indépendants sont utilisés pour l'apprentissage ( $n=100$ ) et le test ( $n=100$ ), ces jeux sont obtenus par photo-interprétation et sont indépendants des échantillons utilisés pour effectuer la sélection des attributs de texture.

Enfin, sur l'image classifiée, une régularisation spatiale est appliquée. Cette régularisation consiste à affecter à chaque pixel, l'étiquette majoritaire dans son voisinage. Le rayon de ce voisinage a été fixé à 4 pixels en résolution MS, ceci correspond à un diamètre de 16 m soit 2 fois l'espacement maximum entre arbres observé *in situ*.

## 4. Résultats et Discussion

### 4.1 Estimation des variables forestières

Les résultats obtenus pour les images Pléiades des mois de Juin et d'Août sont comparables, bien que sensiblement meilleurs en août.

Les meilleurs modèles de régression multiples obtenus pour chaque variable de structure forestière combinent des attributs de texture issus des deux résolutions avec divers paramétrages, pouvant différer d'une date à l'autre. Ceci confirme l'intérêt d'automatiser le processus afin d'avoir une combinaison optimale d'attributs multi-résolutions (Pan et MS) et multi-échelles (différents paramétrages spatiaux).

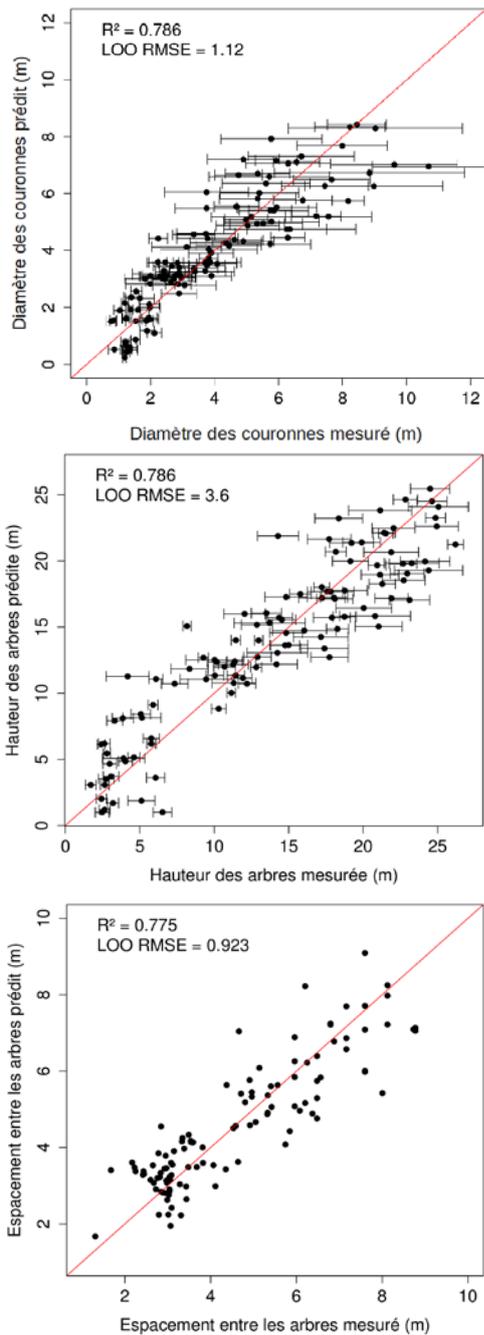
Un résumé des performances de prédiction des variables forestières est présenté dans le Tableau 2. Elles sont similaires pour les deux dates. Les plus faibles performances sur la densité Nah ( $R=0,54$  et  $0,63$ ) traduisent une non-linéarité de sa relation avec les attributs texturaux. Les meilleures précisions d'estimation (LOO RMSE) des autres variables sont les suivantes :  $\sim 1,1$  m pour le diamètre des couronnes Cd (et  $\sim 0,062$  m pour le diamètre des troncs Dbh, très corrélé avec le diamètre des couronnes),  $\sim 3,5$  m pour la hauteur des arbres Ht et  $\sim 0,9$  m pour l'espacement entre les arbres Sp.

Var. Forêt	R <sup>2</sup> multiple		LOO RMSE		Gamme de variation <sup>1</sup>
	Juin	Août	Juin	Août	
Cd (m)	0,79	0,81	1,12	1,10	0,77-10,7
Sp (m)	0,78	0,82	0,92	0,84	1,31-8,77
Ht (m)	0,79	0,83	3,6	3,1	1,7-26,2
Nah (arbres/ha)	0,54	0,63	639	586	150-6729
Dbh (m)	0,83	0,80	0,062	0,066	0,02-0,56

**Tableau 2 :** Récapitulatif des meilleures performances de prédiction des variables de structure forestière obtenues par régressions linéaires multiples. Cd pour le diamètre des couronnes, Sp pour l'espacement entre les arbres, Ht pour la hauteur, Nah pour la densité et Dbh pour le diamètre des troncs.

<sup>1</sup> : Gamme de variation observée *in situ*

La Figure 2 présente les relations entre variables mesurées *in situ* et prédites à partir des meilleurs modèles. Cela pour les trois variables de structure forestière les plus importantes dans la formation de la texture dans les images : Cd, Ht et Sp. La dispersion des résidus varie selon les variables observées. Pour Sp cette distribution semble aléatoire ; pour Ht l'incertitude relative est forte pour des hauteurs inférieures à 5 m. Pour Cd un comportement hétéroscédastique (variance non constante le long de l'échantillon) est observé : plus le peuplement est mature, plus la variabilité locale du diamètre des couronnes est grande et donc plus l'incertitude de la prédiction sera importante. Cet effet est peut-être renforcé par une plus forte imprécision de sa mesure *in situ* pour les arbres les plus hauts (emploi d'une méthode optique). Les bonnes performances sur Cd, avec une RMSE  $\sim 10\%$  de la gamme de variation sur tout le cycle de développement font espérer une identification précise des 5 stades par classification.



**Figure 2 :** Variables forestières mesurées et estimations prédites avec les meilleurs modèles, image de Juin. Les barres correspondent à  $\pm 1$  écart type

#### 4.2 Cartographie des stades de développement

Comme pour l'étape de modélisation, la sélection des attributs de texture par l'importance de variable des RF a conduit à un sous-ensemble d'attributs avec une grande diversité de paramétrages et combinant les deux résolutions. Cette observation est vérifiée dans les étapes de sélection d'attributs pour discriminer la classe C1 des autres et ensuite les quatre classes arborées (C2 à C5) entre elles. Ceci souligne la capacité de notre méthodologie à optimiser l'outil d'analyse de texture ainsi que l'utilisation des données Pléiades en fournissant des descripteurs multi-échelles et multi-résolutions. Les résultats de classification présentés sont ceux obtenus après l'étape de régularisation. Il est

à remarquer que si cette étape apporte une sensible amélioration d'un point de vue visuel, son gain sur les scores globaux de classification n'est pas très important (de l'ordre de 1%).

Le tableau 3 montre la matrice de confusion de la première étape de classification. La forêt en mutation (classe C1) est bien différenciée des autres stades forestiers, cela avec un indice kappa de 0,78 et une précision globale de 94%. On observe tout de même un fort taux d'erreur, aussi bien en omission (22%) qu'en commission (13%) pour cette classe. Ceci est vraisemblablement imputable en partie à la grande variabilité de la végétation à l'intérieur de cette classe.

Référence	Classification			% omission
	C1	Autres	Total	
C1	13313	3787	17100	22
Autres	1984	70016	72000	3
Total	15297	73803	89100	
% commission	13	5		<b>Précision globale = 0,94</b>
				<b>Kappa = 0,78</b>

**Tableau 3 :** Matrice de confusion pour la classification en deux classes : Forêt en mutation (C1) vs Autres classes (C2 à C5), nombre de pixels, image de juin,.

Référence	Classification				Total	% omission
	C2	C3	C4	C5		
C2	14508	1115	474	103	16200	10
C3	1474	15487	2742	97	19800	22
C4	0	774	16919	1207	18900	10
C5	0	969	1026	15105	17100	12
Total	15982	18345	21161	16512	72000	
% commission	9	16	20	9		<b>Précision globale = 0,86</b>
						<b>Kappa = 0,81</b>

**Tableau 4 :** Matrice de confusion pour la classification des stades C2 à C5, nombre de pixels, image de juin.

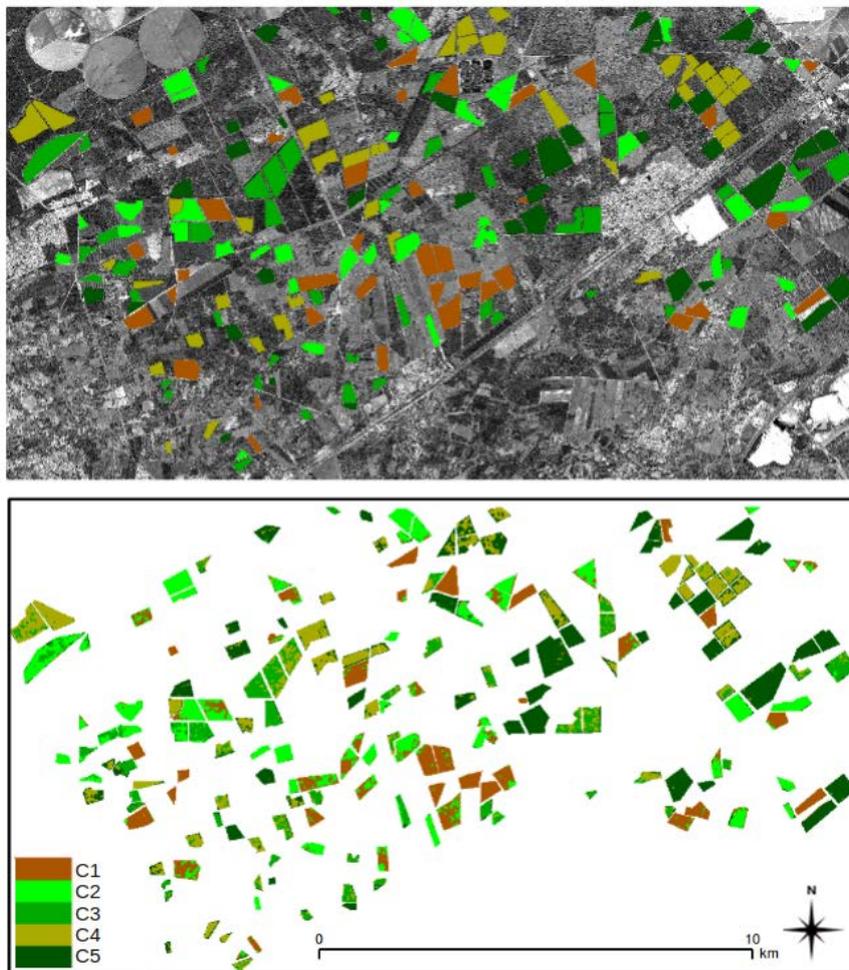
Référence	Classification					Total	% omission
	C1	C2	C3	C4	C5		
C1	13313	3787	0	0	0	17100	22
C2	1983	12657	1020	474	66	16200	22
C3	1	1473	15487	2742	97	19800	22
C4	0	0	774	16919	1207	18900	10
C5	0	0	969	1026	15105	17100	12
Total	15297	17917	18250	21161	16475	81900	
% commiss	13	29	15	20	8		<b>Précision globale = 0,82</b>
							<b>Kappa = 0,78</b>

**Tableau 5 :** Matrice de confusion pour la classification complète des 5 stades par approche hiérarchique, nombre de pixels, image de juin.

Les résultats concernant la seconde étape de discrimination des 4 classes C2 à C5 sont présentés dans le Tableau 4. Les précisions de classification sont satisfaisantes avec un indice kappa de 0,8 et une précision globale de 85%. Les taux d'erreurs les plus importants concernent les classes intermédiaires C3 et C4, ce qui était attendu compte tenu de la nature floue de la frontière entre ces classes puisque la structure forestière présente un continuum. Ceci met en évidence la complexité ce type de cartographie.

Enfin, après combinaison des deux étapes de classification, nous obtenons la matrice de confusion présentée dans le Tableau 5. Le taux de bonne classification globale est de l'ordre de 81%, l'indice Kappa de 0,77. On obtient une amélioration de l'ordre de 10% en utilisant l'approche hiérarchique par rapport

à une classification directe en cinq classes. Ceci a permis d'améliorer considérablement les performances de discrimination des classes arborées C2 à C5. La cartographie produite est donnée en Figure 3 en comparaison avec les classes réelles.



**Figure 3 :** Images des parcelles de référence (sur fond d'image Pan) et de la classification hiérarchique des stades de développement par RF sur le jeu d'images de juin.

## 5. Conclusion

En conclusion, les résultats montrent le fort potentiel de la texture des images Pléiades pour modéliser des variables de structure forestière utiles à l'estimation de la ressource en bois et pour cartographier les stades de développement de peuplements, en particulier sur le massif landais de pin maritime où il a été évalué.

La méthodologie proposée permet d'optimiser l'exploitation de l'information texturale des images Pléiades en calculant des attributs de texture avec une grande diversité de paramétrages possibles et en combinant les deux résolutions panchromatique et multispectrale. Les meilleures combinaisons d'attributs de texture sont ensuite sélectionnées pour estimer les variables de structure forestière et pour la classification des stades développement.

Grâce à l'automatisation des procédures, l'intérêt est de ne faire aucune hypothèse *a priori* sur la relation entre structure forestière et texture d'image, qui peut être très dépendante des conditions d'observation satellitaire

(saison qui détermine la hauteur du soleil et le stade phénologique, angles de visée, par exemple). Ce type d'approche pourrait donc être transposé à d'autres applications et mériterait d'être testé sur des forêts plus complexes ou plus hétérogènes.

La bonne précision obtenue sur l'estimation des variables forestières comme le diamètre des troncs (à attribuer à sa forte relation allométrique à celui des couronnes), la hauteur et l'espacement laisse espérer une application opérationnelle des images Pléiades pour une cartographie fine de la biomasse forestière ou des volumes de bois. Néanmoins la précision sur la hauteur moyenne est moins bonne que celle obtenue sur la hauteur dominante (RMSE ~1 à 2 m) par corrélation d'image sur des observations stéréoscopiques Pléiades par Bélouard *et al.* (2014 dans ce numéro). Un usage en synergie des deux méthodes (analyse de texture et analyse tridimensionnelle sur images THR) viendrait maximiser la précision des estimateurs de la ressource forestière.

Pour la cartographie des stades de développement, une approche hiérarchique est proposée et validée. Une première étape permet de discriminer les peuplements forestiers en mutation (coupes rases, reboisements en cours ou très récents) des autres stades forestiers. Un taux de précision globale de l'ordre de 94% est obtenu. Cette étape est aussi intéressante dans la perspective de la détection de changements (coupe rase, nettoyage ou labour indicateurs d'un reboisement, ou encore dégâts de tempête). La classification en quatre grandes classes allant du stade de jeune peuplement (> 3 ans) à celui de peuplement mature (~50 ans) a atteint un score de précision globale de l'ordre de 85%. Ces résultats, établis en supposant que la localisation de la forêt de pin maritime et les limites de peuplement sont connues, sont très encourageants quant à des perspectives opérationnelles. En effet, à titre de comparaison, la carte des classes d'âge produite occasionnellement par l'IFN (service de l'Inventaire Forestier National) sur le département des Landes (BD Forêt Version 1 de l'IGN) par photointerprétation de prises de vue aérienne (résolution ~50 cm) différencie 3 classes d'âge dans la futaie pure de pin maritime.

La robustesse de la méthodologie proposée et la reproductibilité des hauts niveaux de précision atteints seront à confirmer avec une troisième image Pléiades, acquise dans des conditions très différentes : en hiver (20 Février 2013) avec une hauteur solaire moins élevée et une végétation du sous-bois sèche. L'application aux images multi-annuelles pourra permettre d'évaluer sur le massif landais la possibilité de cartographier des changements tels que des coupes rases, le mitage urbain ou les dégâts de tempête.

## Remerciements

Cette étude a été financée par le Conseil Régional d'Aquitaine et le CNES (Centre National d'Etudes Spatiales). Le CNES a fourni les images Pléiades (© CNES (2012), distribution Airbus DS / Spot Image). Ce travail est fondé sur l'utilisation de l'OrfeoToolBox (Inglada *et al.*, 2009). Les auteurs remercient Bernard Issenhut (INRA, Unité expérimentale Forêt Pierroton) pour la qualité des mesures terrain, Christian Germain (Laboratoire IMS, Bordeaux), Jordi Inglada (CESBIO, Toulouse) et Jean-Pierre Wigneron (INRA, Bordeaux) pour les discussions.

## Références

Beguet B., Guyon D., Boukir S., Chehata N., 2014. *Automated retrieval of forest structure variables based on multi-scale texture analysis of VHR satellite imagery*. ISPRS Journal of Photogrammetry and Remote Sensing, 96: 164–178.

Bélouard T., Py N., Maillet G., Guyon D., Meredieu C., Pausader M., Champion N. *Pinastéréo – Estimation de la hauteur dominante et de la biomasse forestière dans le massif des Landes de Gascogne à partir d'images stéréoscopiques Pléiades*. RFPT, ce numéro (soumis).

Boukir S., Orny C., Chehata N., Guyon D., Wigneron J. P., 2013. *Détection de changements structurels sur des images application en milieu forestier*. Traitement du Signal, 30:401-429.

Breiman L., 2001. *Random forests*. Machine Learning, 45(1):5-32.

Castillo M. A., Ricker M., De Jong B. H. J., 2010. *Estimation of tropical forest structure from spot-5 satellite images*. International Journal of Remote Sensing, 31-10:2767-2782.

Chehata N., Orny C., Boukir S., Guyon D., Wigneron J. P., 2014. *Object based change detection in wind-storm damaged forest using high resolution multispectral images*. International Journal of Remote Sensing, 35(13):4758-4777.

Coueron P., Pelissier R., Nicolini E. A., Paget D., 2005. *Predicting tropical forest stand structure parameters from Fourier transform of very high resolution remotely sensed canopy images*. Journal of Applied Ecology, 42:1121-1128.

Haralick R. M., Shanmugan K., Dinstein I., 1973. *Texture features for image classification*. IEEE Transactions Systems, Man and Cybernetics, 601-621.

Inglada J., Christophe E., 2009. *The Orfeo toolbox remote sensing image processing software*. IEEE International Geoscience and Remote Sensing Symposium, 4:IV-733–IV-736.

Kayitakire F., Hamel C., Defourny P., 2006. *Retrieving forest structure variables based on image texture analysis and ikonos-2 imagery*. Remote Sensing of Environment, 102:390-401.

Maillard P., 2003. *Comparing texture analysis methods through classification*. Photogrammetric Engineering & Remote Sensing, 69(4):357-367.

Maugé J.P., 1987. *Le pin maritime, premier résineux de France*, Editions IDF, Paris, 192p.

Proisy C., Coueron P., Fromard F., 2007. *Predicting and mapping mangrove biomass from canopy grain analysis using fourier-based textural ordination of ikonos images*. Remote Sensing of Environment, 109:379-392.

Regniers O., DaCosta J.-P., Grenier G., Germain C., Bombrun L., 2013. *Texture based image retrieval and classification of very high resolution maritime pine forest images*. Proceedings IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Song C., Dickinson M. B., Su L., Zhang S., Yaussey D., 2010. *Estimating average tree crown size using spatial information from ikonos and quickbird images: Across-sensor and across-site comparisons*. Remote Sensing of Environment, 114:1099-1107.

St-Onge B. A., Cavayas F., 1997. *Automated forest structure mapping from high resolution imagery based on directional semivariogram estimates*. Remote Sensing of Environment, 61:82-95.

Van Coillie F. M., Verbeke L. P., Wulf R. R. D., 2007. *Feature selection by genetic algorithms in object-based classification of ikonos imagery for forest mapping in flanders, belgium*. Remote Sensing of Environment, 110(4):476 -487.