UNIVERSITÉ DES ANTILLES ET DE LA GUYANE

Mémoire d'Habilitation à Diriger des Recherches
(HDR)

Des données brutes raster à l'information vectorielle : processus complet de modélisation des données diffuses

Soutenu le 8 décembre 2011 à l'Université des Antilles et de la Guyane, pôle Guadeloupe

par

Enguerran GRANDCHAMP, Docteur, Maître de Conférences au sein du LAMIA de l'UAG

Devant le jury composé de

Pr. Vincent Charvillat (Examinateur, INPT-ENSEEIHT),
Pr. Christine Fernandez-Maloigne (Rapporteur, Université de Poitiers),
Pr. Kalifa Goïta (Rapporteur, Université de Sherbrooke, Canada),
Pr. Richard Nock (Examinateur, Président de jury, Université des Antilles et de la Guyane),
Pr. Lionel Prevost (Examinateur, Université des Antilles et de la Guyane),
Pr. Florence Sédès (Rapporteur, Université Paul Sabatier)

Résumé

Ce mémoire synthétise mon parcours d'enseignant-chercheur au sein de l'Université des Antilles et de la Guyane de septembre 2002 à septembre 2011. Il retrace mes activités d'encadrement, d'enseignement, de recherche et d'administration en détaillant plus en profondeur mes contributions en recherche. L'application servant de support à l'ensemble de ces contributions est la classification et la représentation des forêts de la Guadeloupe.

Mes thématiques de recherche sont organisées autour d'une chaîne de traitement permettant le passage d'une donnée brute raster vers une information vectorielle. Elle se décompose en trois parties, la première concerne l'extraction d'information à partir d'images satellites et fait successivement appel à la recherche d'espaces couleurs hybrides, à la fusion d'images de résolutions spatiales et spectrales différentes, à l'extraction de descripteurs sur la couleur et la texture des images, la sélection d'attributs et la classification par approche raster. La seconde partie de la chaîne de traitement est consacrée à la modélisation de l'information et à sa représentation sous forme vectorielle. On y retrouve successivement la définition d'un dictionnaire sur les formations forestières, la classification par approche vectorielle, la définition d'un modèle flou vectoriel et la définition d'une couche sémantique. Enfin, la troisième partie de la chaîne concerne l'exploitation de cette information au travers d'un algorithme d'optimisation de l'information et d'un algorithme de coopération raster-vecteur.

Abstract

This report is a synthesis of my experience as lecturer and researcher at the University of French West Indies from September 2002 to September 2011. It details my research, supervision and administrative activities with more details given to my research contributions. The application used as a support for each contribution is the classification and representation of Guadeloupean forests.

My research topics are organized in a framework allowing converting raw raster data into vectorial information. It is composed of three main parts, the first one deals with information extraction from satellite images and successively presents hybrid color spaces choice, image fusion with different spatial and spectral resolutions, color and texture feature extraction, feature selection and classification in a raster way. The second part of the framework is turned to vectorial information modeling and representation. We successively present the definition of a dictionary, a vectorial classification, the definition of a fuzzy vectorial model and the definition of a semantic layer. Finally, the third part of the framework deals with information exploitation through an optimization of the information and raster-vector cooperation algorithm.

Contexte : de la difficulté à la persévérance

Mon parcours d'enseignant-chercheur, loin d'avoir été un long fleuve tranquille, a été très formateur à différents points de vue et mon profil s'est forgé sur le triptyque équilibré d'enseignement, de recherche et d'administration.

Préférant toujours voir le verre à moitié plein qu'à moitié vide, je substitue au *manque de moyens matériels, humains et financiers* à mon arrivée à l'*UAG,* l'opportunité de *monter des projets*, de *tisser un réseau de partenaires*, de *définir une thématique de recherche* et de *gérer un budget*.

J'associe aux *nombreuses ouvertures* (*IUP* 1ère année en 2002, 2ème année en 2003, 3ème année en 2004) et *réorganisations des formations* (passage au *LMD* en 2006 et refonte de la maquette en 2010) ainsi qu'à la participation depuis 2007 à la *création d'une école d'ingénieur* (dossier en cours d'expertise à la *CTI*), l'opportunité de *créer des enseignements*, d'*élargir mon champ disciplinaire*, de *réfléchir à la finalité d'une formation* et de *viser l'excellence*.

Je vois en une *petite structure* avec un *turnover important*, l'*enrichissement des relations humaines* et la nécessité *d'une stabilité transversale*.

J'interprète l'*isolement géographique* et l'*insularité* de mon laboratoire comme un facteur de *cohérence avec les problématiques locales* me permettant de *donner une identité à mes travaux*.

Enfin, je remplace aisément les *charges administratives* liées aux *différentes responsabilités* au sein du laboratoire et pendant sa restructuration (*GRIMAAG* 2002-2010, *LAMIA* 2010-…) par la satisfaction de participer à l'*animation de la vie collective*.

Le leitmotiv de ce parcours est une constante activité de recherche, avec toute la modestie qu'on peut lui attribuer, me donnant une ouverture scientifique sur plusieurs communautés (télédétection, *SIG*, modélisation de données) et me permettant de consolider mes contributions en les valorisant notamment grâce à ce mémoire qui en est l'un des premiers aboutissements appuyé par des encouragements nationaux, au travers de la *PEDR* en 2004 et d'une évaluation de *rang A* du ministère en 2009, et des encouragements locaux au travers de la *PES* en 2009 et d'un *CRCT* en 2011.

Et pour toutes les choses qui ne peuvent être exprimées ici …

"*Don't let the noise of others' opinions drown out your own inner voice.*
And most important, have the courage to follow your heart and intuition.
They somehow already know what you truly want to become.
Everything else is secondary."

"*Ne laissez pas le brouhaha extérieur étouffer votre voix intérieure.*
Ayez le courage de suivre votre cœur et votre intuition.
L'un et l'autre savent ce que vous voulez réellement devenir.
Tout le reste est secondaire."

Extrait et traduction du discours de *Steve Jobs,* disparu le 5 Octobre 2011, pour la cérémonie de remise de diplôme de l'Université de Stanford le 12 Juin 2005 [**121**].

Table des matières

1 LISTE DES FIGURES ET DES TABLEAUX 6

2 GLOSSAIRE DES TERMES ET DES ABREVIATIONS 7

2.1 ABRÉVIATIONS 7
2.2 TERMINOLOGIE 8

3 NOTICE INDIVIDUELLE 9

3.1 ETAT CIVIL 9
3.2 FORMATION 9
3.3 QUELQUES DATES IMPORTANTES 9
3.4 PARCOURS ET THÉMATIQUES D'ENSEIGNEMENT 10
3.5 PARCOURS ET THÉMATIQUES DE RECHERCHE 11
3.6 ENCADREMENTS 12
3.7 PROJETS ET COLLABORATIONS 18
3.8 PUBLICATIONS 18
3.9 RESPONSABILITÉS ET FONCTIONS ÉLECTIVES 21
3.10 ANIMATION SCIENTIFIQUE 22
3.11 PRIMES DE RECHERCHE 22
3.12 SYNTHÈSE CHRONOLOGIQUE 23

4 INTRODUCTION 24

4.1 PARCOURS EN RECHERCHE 24
4.2 SYNTHÈSE DES CONTRIBUTIONS 26
4.3 ORGANISATION DU MÉMOIRE 28

5 CONTRIBUTIONS 29

5.1 LE CONTEXTE 30
5.1.1 LES SCIENCES D'INFORMATION GÉOGRAPHIQUES (SIG) 30
5.1.2 DIFFÉRENTES REPRÉSENTATIONS D'UNE MÊME RÉALITÉ 32
5.1.3 QUELQUES VERROUS 34
5.1.4 L'OPTIMISATION EN TOILE DE FOND 35
5.1.5 BILANS 36
5.2 PRÉPARATION DES DONNÉES ET EXTRACTION D'INFORMATION 37
5.2.1 TRAITEMENT D'IMAGE 37
5.2.2 EXTRACTION DE CARACTÉRISTIQUES 42
5.2.3 SÉLECTION D'ATTRIBUTS 47
5.2.4 CLASSIFICATION D'IMAGES SATELLITES 54
5.3 MODÉLISATION DES DONNÉES 57
5.3.1 LE DICTIONNAIRE DES ESPACES FORESTIERS DE LA GUADELOUPE 58
5.3.2 CLASSIFICATION PAR APPROCHE VECTEUR 60
5.3.3 ADAPTATION DES FORMATS VECTORIELS POUR LES DONNÉES DIFFUSES 65
5.3.4 MODÉLISATION SÉMANTIQUE DES SCÈNES 72
5.4 EXPLOITATION DES DONNÉES 75

5.4.1 OPTIMISATION DE LA SÉLECTION D'INFORMATION 75
5.4.2 COOPÉRATION RASTER-VECTEUR 78
5.5 SYNTHÈSE DE L'ORGANISATION DES THÉMATIQUES 80

6 PERSPECTIVES 82

6.1 VERS UNE COHÉSION AUTOUR DES *SIG* 82
6.2 PERSPECTIVES À COURT ET MOYEN TERMES 83
6.2.1 LA SÉLECTION D'INFORMATION 83
6.2.2 AFFINER LA CLASSIFICATION VECTORIELLE DES FORÊTS 85
6.2.3 VERS UNE CLASSIFICATION 100% OBJET 86
6.3 PERSPECTIVES À LONG TERME 87
6.3.1 STRICT MAIS IMPRÉCIS, DIFFUS MAIS PRÉCIS : VERS UN MÊME MODÈLE ? 87
6.3.2 COUCHE SÉMANTIQUE EXTERNE OU SÉMANTIQUE INTÉGRÉE ? 91
6.3.3 DE LA NÉCESSITÉ D'UNE STANDARDISATION À L'UTILISATION D'UN FORMALISME : DES OUTILS ORIENTÉS UTILISATEUR 92
6.3.4 LES RELATIONS ENTRE OBJETS ET CONCEPTS 94
6.4 SCHÉMA SYNTHÉTIQUE 96

***7* CONCLUSION: UNE PROJECTION AU MILIEU DES DEFIS DES *SIG* 98**

8 ANNEXES 99

8.1 RAPPORTS 99

9 REFERENCES 100

1 Liste des figures et des tableaux

Tableau 1 - Enseignements **10**
Tableau 2 - comparaison de l'indice de qualité (GQ) des différentes méthodes de fusion sur les composantes *R, V* et *B*. **41**
Tableau 3 - Comparaison de l'approche multi-objectif avec d'autres approches **51**
Tableau 4 - Taux de bonne classification et stabilité pour quelques sous-ensembles intéressants **53**
Tableau 5 - Taux de bonne classification et Erreur Globale pour les données de synthèse **55**
Tableau 6 - Matrice de confusion entre *FT* et la classification de référence (% de surface) **63**
Tableau 7 - Classes d'habitat **64**
Tableau 8 - Hiérarchie des transitions **71**

Figure 1 - Schéma synthétique simplifié de la chaîne de traitement **29**
Figure 2 - Raster vs Vecteur (extrait de [117]) **32**
Figure 3 – Ensemble de Pareto pour deux critères **36**
Figure 4 – Projection de différentes classes dans différents espaces couleurs (a) IST (b) LAB (c) HSV (d) RVB **37**
Figure 5 – Comparaison des algorithmes de recherche d'un espace couleur hybride **39**
Figure 6 – Comparaison des espaces hybrides **40**
Figure 7 - Comparaison de résultats de fusion d'images entre *TH et le modèle 1* **42**
Figure 8 - Exemple de caractérisation **44**
Figure 9 - Extraits d'images représentant des parcelles agricoles (A_i) et des forêts (F_i) et Projection des moyennes des couleurs sur les composantes *S* et *T* **44**
Figure 10 - Textures et histogrammes correspondants **46**
Figure 11 – Histogramme des Moments de Hu calculés sur chaque texture **46**
Figure 12 - Dimension fractale **46**
Figure 13 - Spectre de Legendre **47**
Figure 14 - Comparaison des stabilités pour différentes bases (a) wine (b) imgSeg (c) ionosphere (d) landSat **52**
Figure 15 - Analyse de la stabilité pour la base landSat : (a) Front de Pareto (b) Sous ensembles dominants **53**
Figure 16 - Exemple de données de synthèse pour la validation des classifieurs (a) image basse résolution, (b) image haute résolution (c) répartition des classes **55**
Figure 17 - Exemple de données réelles (classées par des experts) pour la validation des classifieurs **55**
Figure 18 – Résultats de la classification des images de synthèse **55**
Figure 19 - Résultats de la classification des images naturelles **56**
Figure 20 - Séparation forêt agriculture sur des images Spot 5 ..... **56**
Figure 21 - Séparation forêt agriculture sur une scène Quickbird **57**
Figure 22 - Ontologie des couverts forestiers de la Guadeloupe **59**
Figure 23 - Extraits de la base de textures **60**
Figure 24 - Localisation des cimes et des couronnes des arbres **60**
Figure 25 - Cartographie des unités écologiques naturelles de la Guadeloupe. (La légende ne référence que les unités représentées en Basse-Terre [181]) **61**
Figure 26 – (a) Données d'apprentissage (Placettes d'observation) et (b) Extrait de la carte des unités. **62**
Figure 27 – (a) Carte Ecologique (1996) - Classification supervisée (b) C4.5 et (c) FT (2011) **62**
Figure 28 – Parc Naturel de la Guadeloupe : (a) Carte Ecologique (1996) (b) Classification supervisée FT (2011) (c) Différences (noir) **63**
Figure 29 - Classification des habitats en 1996, 2000 et 2004 avec les méthodes d'apprentissage FT et C **65**
Figure 30 - Cartes Floues de 9 formations forestières **68**
Figure 31 - Différentes représentations vectorielles simplifiées d'une formation forestière (classe 9) **69**
Figure 32 - *Classes de transition* **71**
Figure 33 - Différentes représentations d'une même entité **72**
Figure 34 - Liaison couche thématique et ontologie **73**
Figure 35 - Résolution de conflits sur des données simulées **74**
Figure 36 - Différents arbres pour un même ensemble de concepts **76**
Figure 37 - Agrégation hiérarchique **77**
Figure 38 - Résultat de la fusion d'information **77**

Figure 39 - Classification raster des unités écologiques vectorielles **79**
Figure 40 - Extraction d'unités écologiques vectorielles à partir d'une classification raster **79**
Figure 41 - Schéma synthétique complet de la chaîne de traitement **81**
Figure 42 – Compacité des concepts d'une couche ($\delta = 2$) **84**
Figure 43 – Critères de l'approche par voisinage pour la sélection d'information **85**
Figure 44 – Tolérance de positionnement **89**
Figure 45 – Représentation du gradient de la fonction d'appartenance **90**
Figure 46 – Polygone ouvert (extrait de [82]) **91**
Figure 47 – (a) Couche sémantique externe - (b) Couche augmentée **92**
Figure 48 – Construction du graph d'adjacence valué **95**
Figure 49 – Synthèse des contributions et perspectives **97**
Figure 50 – Répartition thématique des auteurs basée sur un extrait de [108] **99**

2 Glossaire des termes et des abréviations

2.1 Abréviations

2OMF	***2 O**bjectifs **M**ulti **F**ront*
AG	**A**lgorithmes **G**énétiques
BRGM	**B**ureau de **R**echerches **G**éologiques et **M**inières
CESAR	**C**lassification d'**ES**pèces **Ar**borescentes
CESBIO	**C**entre d'**E**tudes **S**patiales de la **BIO**sphère
CIRAD	**C**entre de coopération **I**nternationale en **R**echerche **A**gronomique pour le **D**éveloppement
CNES	**C**entre **N**ational d'**E**tude **S**patiale
CNU	**C**onseil **N**ational des **U**niversités
Corr	**C**orrélation
CRCT	**C**ongé pour **R**echerche ou de **C**onversion **T**hématique
CTI	**C**ommission des **T**itres d'**I**ngénieur
D	Pertinence
DAF	**D**irection de l'**A**griculture et des **F**orêts
DDB	**D**éveloppement **D**urable et **B**iodiversité
DFA	**D**épartements **F**rançais d'**A**mérique
DYNECAR	Laboratoire de **DYN**amique des **E**cosystèmes **CAR**aïbes
ENSEEIHT	**E**cole **N**ationale **S**upérieure en **E**lectronique **E**lectrotechnique **I**nformatique **H**ydraulique et **T**élécommunications
FSDD	**F**eature **S**election algorithm based on a **D**istance **D**iscriminant
GKD	**G**eographic **K**nowledge **D**iscovery
GPS	**G**lobal **P**ositioning **S**ystem
GRIMAAG	**G**roupe de **R**echerche en **I**nformatique et **M**athématiques **A**ppliquées des **A**ntilles et de la **G**uyane
IC	**C**ritères d'**I**nformation
IM	**I**nformation **M**utuelle
INPT	**I**nstitut **N**ational **P**olytechnique de **T**oulouse
IRD	**I**nstitut de **R**echerche de **D**éveloppement
IRIT	**I**nstitut de **R**echerche en **I**nformatique de **T**oulouse

IST	**I**ntensité **S**aturation **T**einte
IUP	**I**nstitut **U**niversitaire **P**rofessionnalisé
KDD	**K**nowledge **D**iscovery from **D**ata
KNN	**K**-**N**earest **N**eighbor
LAMIA	**LA**boratoire de **M**athématiques et **I**nformatique **A**ppliqués.
LDA	**L**inear **D**iscriminant **A**nalysis
LIMA	**L**aboratoire d'**I**nformatique et de **M**athématique **A**ppliquée
LMD	**L**icence **M**aster **D**octorat
MAH	**M**ahalanobis
MAI	**M**éthodes d'**A**xes **I**ndépendants
MMG	**M**odèle de **M**élange de **G**aussiennes
MP	**M**éthodes **P**erceptuelles
mRMR	***m**in **R**edundancy **M**ax **R**elevance*
mRMR-PC	***m**in **R**edundancy **M**ax **R**elevance **P**areto **C**urve*
NB	**N**aive **B**ayes
OGC	**O**pen **G**eospatial **C**onsortium
ONF	**O**ffice **N**ational des **F**orêts
OWL	**O**ntology **W**eb **L**anguage
PARAGE	occu**P**ation **A**gricole dans les **R**égions **A**ntilles et **G**uyane
PAS	**P**lan d'**A**ction **S**tratégique
PD	**P**ouvoir **D**iscriminant
PEDR	**P**rime d'**E**ncadrement **D**octoral et de **R**echerche
PES	**P**rime d'**E**xcellence **S**cientifique
PNN	**P**robabilistic **N**eural **N**etwork
PVM	**P**redictive **V**egetation **M**odeling
R	**R**edondance
RDF	**R**esource **D**escription **F**ramework
SI	**S**ystème d'**I**nformation
SIG	**S**ystème ou **S**cience d'**I**nformation **G**éographique
SVM	**S**upport **V**ector **M**achine
SIGMA	**S**ystème d'**I**nformation **G**éographique de la **MA**rtinique
TH	**T**ransformation **H**ybride
UAG	**U**niversité des **A**ntilles et de la **G**uyane
UE	**U**nité **E**cologique
UER	**U**nité **E**cologique **R**aster
UEV	**U**nité **E**cologique **V**ectorielle
XML	e**X**tensible **M**arkup **L**anguage

2.2 Terminologie

Donnée	Donnée brute sans information contextuelle ou sémantique
Information	Donnée annotée d'une information contextuelle et/ou sémantique
Objet (Vecteur)	Donnée stockée ou représentée sous forme vectorielle
Raster	Donnée stockée ou représentée sous forme de grille (image, carte scalaire, etc.)

3 Notice individuelle

3.1 Etat Civil

Enguerran GRANDCHAMP,
Né le 25 août 1975 à Paris (14ème)
Nationalité française

12 rue Th. Gendrey
97139 ABYMES

Maître de Conférences en 27ème section à l'Université des Antilles et de la Guyane (UAG)
Laboratoire de Mathématiques et d'Informatique Appliqués (LAMIA)
Département Mathématiques et Informatique (DMI).
Campus de Fouillole, 97157 Pointe-à-Pitre Cedex, BP 250
Tél : 05.90.48.30.75 Fax : 05.90.48.30.86 Email : enguerran.grandchamp@univ-ag.fr

3.2 Formation

- ✓ Depuis septembre 2002 : **Maître de Conférences en 27ème section** à l'Université Antilles Guyane (UAG)
- ✓ 2001-2002 : **ATER à l'INPT-ENSEEIHT à Toulouse**
  Département : Informatique et Mathématiques Appliquées
  Laboratoire de recherche : IRIT-LIMA Collaborateur : Alcatel Space Industries (Toulouse)
- ✓ 1998-2001 : **Doctorant allocataire de recherche en informatique et moniteur**
  Titre de la thèse **: quelques contributions pour l'optimisation de constellations de satellites**
  Directeur : Pr. J. Noailles
  Jury : Pr. P. Alliot, Pr. D. Brousse, Pr. V. Charvillat, Dr. P. Dago, Pr. J.K. Hao, Pr. J. Noailles
  Soutenance à l'ENSEEIHT le 17 décembre 2001 (mention très honorable)
  Laboratoire de recherche : IRIT-LIMA Collaborateur : Alcatel Space Industries
- ✓ 1997-1998 : **DEA d'Informatique de l'Image et du Langage** (mention bien)
  Titre : **Segmentation d'images radars et détection de ruptures par ondelettes**
  Laboratoire de recherche : IRIT-LIMA Collaborateurs : CNES-CESBIO
- ✓ 1995-1998 : **Diplôme d'ingénieur en Informatique et Mathématiques Appliquées** INPT-ENSEEIHT

3.3 Quelques dates importantes

2001 Diplôme de doctorat
2002 Maître de Conférences à l'UAG
2004 Bénéficiaire de la Prime d'Encadrement Doctoral et de Recherche (PEDR)
2009 Bénéficiaire de la Prime d'Excellence Scientifique (PES)
2011 Obtention d'un Congé pour Recherche ou Conversion Thématique (CRCT)

3.4 Parcours et thématiques d'enseignement

Après avoir obtenu mon diplôme d'ingénieur en Informatique et Mathématiques Appliquées de l'INPT-ENSEEIHT et mon DEA de l'Université Paul Sabatier (1998), j'ai commencé à enseigner en tant que moniteur durant ma thèse au sein de l'INPT. J'ai eu l'occasion de rencontrer divers publics au cours des 3 années de monitorat et d'une année d'ATER : élèves ingénieurs en informatique ou électronique en formation initiale, ingénieurs en spécialisation ou en formation continue, élèves ingénieurs polytechniciens, etc. J'ai également commencé à créer des enseignements lors de l'ouverture de la filière télécom-réseau de l'ENSEEIHT avec notamment un Bureau d'Etude sur la triangulation GPS faisant appel à des techniques d'optimisation.

A mon arrivée à l'UAG, l'ouverture des trois années de l'IUP I2M échelonnée entre 2002 et 2004 a nécessité de créer (support de cours, sujets de TP, TD, etc.) un certain nombre d'enseignements dans des domaines proches de mes thématiques de recherche (Image, Optimisation …), proches des enseignements que j'avais déjà dispensé (Programmation Orientée Objets, UML) ou très éloignés (Réseaux, Systèmes Répartis, XML). Avec le passage à l'organisation LMD en 2006, l'arrivée de nouveaux enseignants et l'ouverture du Master d'Informatique mes enseignements ont à nouveau changé et se sont recentré principalement sur la conception orientée objet (Java et UML en Licence et Master, projet pluridisciplinaire faisant intervenir la programmation objet, les systèmes répartis et les bases de données) ainsi que sur le développement de l'enseignement des SIG dans différentes filières (informatique, géologie, biologie). D'autres enseignements plus théoriques ou techniques tels que la complexité, l'optimisation combinatoire ou l'analyse d'image complètent le panel de mes enseignements.

J'ai créé et enseigné en Cours, TD et TP dans l'ensemble des enseignements présentés dans le tableau suivant.

Total : Licence : 1279 h, Master : 940 h

Enseignement	Total	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Total	2219	271	221	227	209	239	233	246	265	96	210
Java UML	662	64	64	45	29	78	75	86	83	45	93
Réseaux	164	41	41	41	41						
Réseaux	164	41	41	41	41						
Image	191	33	35	35	35			30	23		
Algo. prog.	193	72	35	35		33	18				
Optimisation	69			25	25	19					
Complexité	270				35	45	45	45	45	15	40
UML	58					29	29				
XML	82					19	21	19	23		
Projet pluri.	104					16	16	16	20	10	26
Arch. Rép.	92						29	29	34		
Multimédia	46							21	25		
SIG	64								12	26	26
SIG	25										25

Tableau 1 - Enseignements

3.5 Parcours et thématiques de recherche

Mon parcours en recherche depuis mon DEA démarré en 1997 m'a conduit à aborder plusieurs thématiques me permettant de couvrir des domaines tels que la télédétection, l'analyse d'images, la classification, l'optimisation combinatoire, la modélisation de données et les Systèmes d'Information Géographiques.
Ces différentes thématiques ont été abordées de manière indépendante durant

1. mon DEA de 1997-1998 : *Segmentation d'images radars et détection de ruptures par ondelettes : application à la détection de cible* (décomposition en ondelettes, suivi de chaînes de maxima, etc.),
2. ma thèse de doctorat de 1998-2001 : *Optimisation de constellations de satellites* (modélisation de systèmes complexes, optimisation combinatoire par approche classique et heuristique)
3. et mes premières années en tant que Maître de Conférences de 2002-2008 : *Aide à la segmentation d'images satellites par caractérisation de ruptures, Classification d'espèces arborescentes par analyse de couleur et de textures d'images satellites.*

C'est en 2008 que j'oriente mes recherches vers les **Systèmes d'Information Géographique** (*SIG*). Plus qu'un véritable changement thématique, ce contexte scientifique et technologique offre la possibilité de combiner et mettre en œuvre l'ensemble des outils et techniques que j'ai abordés jusque là autour de problématiques liées à la valorisation et à la protection des ressources forestières. En effet, l'association d'outils d'analyse spatiale, de géostatistique ou de systèmes d'information donne une autre dimension aux traitements jusque là cloisonnés. Ceci a permis de définir une chaîne de traitement complète permettant le passage d'une donnée brute image dite *raster* (image satellite principalement) à une information exploitable *objet* (représentée de manière vectorielle sous forme de couches d'information). Le passage de l'une à l'autre nécessite de traiter, analyser et combiner différentes sources de données. Ceci fait successivement appel à du **traitement d'image** (construction d'espaces couleurs hybrides, fusion d'images de résolution spectrale et spatiale différentes), de l'**analyse d'image** (analyse de textures, de couleurs, avec des approches géométrique, structurelle, statistique, fractale, etc.), de la **classification** (supervisée ou non supervisée), de la **modélisation de données** par ensembles flous, de la **fusion d'informations** (croisement de couches d'information vectorielles et raster, correction d'erreurs, relations spatiales, ontologies) et de l'**optimisation** (sélection optimale d'attributs et d'information).

Enfin, la valorisation des ressources forestières a fait émerger des problématiques de modélisation de données et d'analyse sémantique non encore résolues au sein des SIG avec notamment le problème de la représentation de données aux frontières diffuses dans un format vectoriel objet. Ces problématiques posent le problème plus général de l'adaptation des structures de données à la modélisation de certains phénomènes ainsi que celui de la résolution des conflits entre différentes représentation d'une même réalité. Ces deux derniers points définissent l'axe principal de mes recherches à venir. Ces différentes thématiques ont été abordées et le sont encore au travers de thèses, stages de Master, projets et recherches personnelles et ont donné lieu à des publications et collaborations locales (laboratoire Dynecar, Parc National, ONF), nationales (laboratoire XLIM-SIC de l'Université de Poitiers) et internationales (Université de Curitiba au Brésil et de Moncton au Canada).

3.6 Encadrements

Doctorat

Laurent Manyri (octobre 2002 à décembre 2005)

Titre : *Développement des capteurs logiciels pour la caractérisation des populations microbiennes. Application aux procédés de fermentation : énergies renouvelable, boissons alcoolisées*

- ✓ Thèse soutenue le 14 décembre 2005 à l'INSA de Toulouse [**148**]
- ✓ Directeur de thèse : Pr. J Desachy (UAG), Co-Directeur : Dr. A Donscescu (LAAS)
- ✓ Jury : Pr. JP Asselin de Beauville, Pr. I. Bloch, Pr. M. Cheriet, Dr. A Donscescu, Pr. J Desachy, Pr. G. Gomat, Dr. **E. Grandchamp**,
- ✓ Situation actuelle : Maître de Conférences au sein du département Gestion Logistique et Transport de l'IUT de Kourou depuis 2005.
- ✓ Ma contribution
    - Taux d'encadrement : 30%
    - Aide technique sur les méthodes de segmentation et de traitement d'images.
- ✓ Détails des travaux : ces travaux concernent le suivi de la production de ferments alcooliques. Les techniques habituelles de contrôle sont basées sur des prélèvements longs à réaliser et des analyses destructives des ferments. La méthode alternative proposée ici est d'utiliser un microscope in-situ permettant d'acquérir des images régulières des bacs à fermentation sans destruction. Des algorithmes de segmentation d'images et de reconnaissance des formes sont ensuite appliqués afin de compter les bactéries, de les classer par type et de détecter leur stade d'évolution (bourgeonnement, maturité, ...). Ce suivi permet de décider des traitements à appliquer sur les bacs (injection de produits, changement de température, poursuite ou arrêt de la fermentation, etc.). Les champs d'application sont nombreux et inclus la production d'énergies renouvelables, de biocarburants et de boissons alcoolisées.
- ✓ Production : co-dépôt du **brevet ANALYEAST** porté par le LAAS UPR 8001 [**34**]

Mohamed Abadi (octobre 2004 à juin 2008)

Titre : *Couleur et texture pour la représentation et la classification d'images satellites multi résolutions*

- ✓ Thèse soutenue le 30 juin 2008 à l'UAG [**54**]
- ✓ Mention très honorable avec proposition par le jury pour le prix de thèse de l'UAG
- ✓ Directeur de thèse : Pr J. Desachy Co-Directeur : Dr. **E. Grandchamp**
- ✓ Jury : Dr. HDR P. Couteron, Pr. J. Desachy, Dr. HDR P. Gançarski, Dr. **E. Grandchamp**, Pr. M. Herbin, Pr. R. Nock, Pr. J. Vaillant, Cr. HDR H. Yahia
- ✓ Situation actuelle : ingénieur de recherche à l'Université de Poitiers.
- ✓ Ma contribution
    - Encadrement : 95 %
    - Définition du sujet de la thèse

- Recherche de financement pour la thèse (rémunération, missions, matériel)

✓ Détail des travaux : cette thèse s'est déroulée dans le cadre du projet CESAR (Classification d'Espèces Arborescentes). Ce projet a été cofinancé dans le cadre du programme européen INTERREG IIIb Espace Caraïbes à hauteur de 84 000€ pour l'UAG (budget global de 237 000 €). Ce projet regroupait trois partenaires basés en Guadeloupe, au Brésil et au Canada. L'objectif était le recensement de caractéristiques de forêts par télédétection. Trois axes de recherche ont été définis : la classification par analyse de textures pour le pôle Guadeloupe, la fusion de données hétérogènes pour le pôle Brésil et l'extraction de mesures pour le pôle Canada. L'utilisation des images satellites (IKONOS, QuickBird) à très haute résolution spatiale (moins de 1m) permet d'envisager un recensement à l'arbre près et ainsi permettre une évaluation fine des ressources forestières et un contrôle fiable par les autorités de gestion. Les techniques utilisées pour la classification font appel à la fusion d'images multi-résolutions, à la modélisation des couleurs et des textures par analyse fréquentielle, structurelle, fractale et multi fractale, avec notamment la recherche d'un espace couleur hybride optimal. Par ailleurs, une participation à la première année de thèse de Lucas Luis Alberto sur le pôle Brésil a également été menée (définition du sujet, recherche bibliographique).

✓ Production : 10 publications sont liées au projet CESAR [**2**], [**5**], [**16**] à [**20**], [**31**], [**32**], [**43**].

Saliha Loumi et Farid Alilat (janvier à avril 2007)

Titre : *Caractérisation de textures couleurs*

✓ Situation actuelle : enseignants-chercheurs à Alger

✓ Ma contribution

- Encadrement : 100% pendant 4 mois
- Définition du sujet et du protocole de comparaison

✓ Détail des travaux : ces travaux se sont déroulés dans le cadre d'un stage de doctorat d'état de l'Université d'Alger. L'objectif était de comparer deux approches pour la caractérisation des textures : l'une basée sur l'approche fractale, l'autre sur les réseaux de neurones.

✓ Production : une soumission dans la revue internationale *Transaction on Geoscience and Remote Sensing* (TGRS) [**37**].

Artur José Freire Gil (avril 2008 à juin 2008)

Titre : *Fusion des descripteurs couleurs et textures pour la classification d'images forestières*

✓ Situation actuelle : post doctorant aux Etats-Unis

✓ Ma contribution

- Encadrement à 100% pendant 3 mois.
- Compléments de formation sur la télédétection.

✓ Détail des travaux : ces travaux se sont déroulés dans le cadre d'un stage de doctorat de première année de l'Université des Açores. La thèse porte sur la classification des zones forestières des Açores par une approche SIG. Après une période de formation, le travail a porté sur l'utilisation conjointe de descripteurs de couleurs et de textures pour caractériser les

différents types de couverts forestiers. Ce travail lui a permis d'appréhender les difficultés du traitement d'images et son intérêt pour compléter son approche.

Laurent Girdary (depuis octobre 2010)

Titre : *Analyse spatiale de l'habitat de Guadeloupe pour l'étude des facteurs de propagation du virus de la dengue*

- ✓ Directeur : L. Marrama — Co-directeur : **E. Grandchamp**
- ✓ Situation actuelle : étudiant en dernière année de thèse, en cours de rédaction.
- ✓ Ma contribution : encadrement à 100% depuis octobre 2010.
- ✓ Détail des travaux : l'encadrement de la fin de thèse a pour objectif une étude comparative de la classification supervisée et non supervisée des habitats de Guadeloupe dans un contexte raster et objet (vectoriel). Les techniques utilisées font appel à des arbres de décision basés sur la fusion d'informations hétérogènes (données démographiques, physiques et spatiales).
- ✓ Production : un article soumis en juillet 2011 dans la revue *International Journal of Geographical Systems* [**38**].

Wilfried Segretier (depuis octobre 2010)

Titre : *Sélection optimale de l'information par combinaison de couches d'information vectorielle et raster dans un SIG par approche heuristique*

- ✓ Directeur de thèse : M. Collard — Co-directeur : **E. Grandchamp**
- ✓ Situation actuelle : en deuxième année de thèse
- ✓ Ma contribution
  - encadrement à 50%
  - définition du sujet, de la modélisation du problème et de l'approche algorithmique
  - développement de la partie serveur et interfaçage des SIG
- ✓ Détail des travaux : la fusion de multiples couches d'informations génère souvent une information trop morcelée pour être exploitable. Il convient alors de regrouper certaines entités pour rendre plus lisible et interprétable l'information. Le respect de la sémantique des données permet de donner un sens aux regroupements en définissant des concepts généraux au sein d'une ontologie. La forte combinatoire du problème posé nous a conduits à adopter une approche par optimisation approchée basée *(i)* sur des algorithmes génétiques à partir de plusieurs codages de l'information faisant intervenir des ontologies *(ii)* sur une approche par voisinage afin de comparer les performances (qualité des résultats et temps de calcul). Les deux approches sont réalisées dans un contexte multi-objectif (qualité et quantité d'information représentée) afin de fournir à l'utilisateur un ensemble Pareto optimal de regroupements.
- ✓ Production : un article accepté à la conférence internationale Geoprocessing'2012 [**8**] et un en cours de soumission à la conférence SAGEO'2012 [**41**].

Evelin Fonseca Cruz (depuis octobre 2010)

Titre : *Diagnostic et renforcement d'un réseau électrique : étude de l'impact cyclonique sur les réseaux électriques*

- ✓ Directeur de Thèse : Didier Puzenat Co-directeur : **E. Grandchamp**
- ✓ Situation actuelle : en deuxième année de thèse en cotutelle avec l'Université de La Havane à Cuba.
- ✓ Ma contribution
  - encadrement à 50%
  - recherche de données historiques d'apprentissage et de validation sur la Guadeloupe (partenaire EDF)
  - intégration des données au sein des SIG
- ✓ Détail des travaux : faisant suite à la thèse de Mathias Peroumalnaik soutenue en décembre 2010 à l'UAG sur l'étude de l'impact cyclonique sur le bâti, cette thèse vise à étendre et adapter les outils mis en place afin de les appliquer sur les réseaux électriques. L'objectif à terme est de fournir aux décideurs un outil d'aide au diagnostique d'un réseau électrique afin de planifier les interventions pour le renforcer. Les outils utilisés sont les systèmes de classeurs, les SIG, les réseaux de neurones.

Stages de Master recherche 2ème année (ou assimilé)

Mathieu Convert (janvier 2004 à juillet 2004)

Titre : *développement d'une base logicielle de traitement d'images*

- ✓ Stage de fin d'étude d'ingénieur en informatique de l'EPITA.
- ✓ Encadrement : 100 %
- ✓ Situation actuelle : ingénieur en informatique
- ✓ Détails : L'objectif de ce stage était de développer une base logicielle permettant d'harmoniser mes développements en traitement d'image. La modélisation et la conception ont été réalisées de manière à autoriser l'intégration de tout type de données et d'algorithmes. Elle permet de centraliser les développements pour faciliter la réutilisation et la présentation des recherches. Elle fournit également un certain nombre de modules (affichage, importation, exportation, paramétrage, ….) qui évitent des développements non directement liés aux recherches.
- ✓ Production : base logicielle de traitement d'images

Guianni Commin (février à juillet 2004)

Titre : *Reconnaissance de Textures de forêts*

- ✓ Stage de DEA
- ✓ Encadrement : 100 %
- ✓ Détails : cette étude est un travail préliminaire d'expérimentation de différentes techniques de

caractérisation de textures (matrices de cooccurrence, méthodes statistiques, ondelettes, etc.).

- ✓ Production : guide technique sur les méthodes de caractérisation

Ikram El Missi (janvier à juillet 2007)

Titre : *Suivi du front forêt-agriculture par analyse de textures*

- ✓ Stage de fin d'étude
- ✓ Encadrement : 50 %
- ✓ Situation actuelle : Ingénieure en informatique en Ile de France.
- ✓ Détails : suite aux résultats obtenus dans le cadre du projet CESAR, une collaboration a été mise en place dans le cadre du projet PARAGE (occu**P**ation **A**gricole dans les **R**égions **A**ntilles et **G**uyane) porté par SpotImage, le CIRAD, l'IRD et SIGBEA. L'objectif du stage était le suivi du front forêt-agriculture par analyse de textures. Les applications en découlant portent sur le suivi des exploitations agricoles. Le travail a consisté en une déclinaison d'outils développés dans le cadre du projet CESAR.
- ✓ Production : intégration des outils dans le démonstrateur développé par SIGBEA et destiné aux collectivités locales et aux industriels (http://parage.sigbea.fr) [**45**]

Frank Duhamel (janvier à juillet 2008)

Titre : *Obtention d'unités écologiques homogènes par relaxation de couches d'un SIG*

- ✓ Stage de fin d'étude
- ✓ Encadrement : 100 %
- ✓ Détails : les unités écologiques sont définies par les biologistes comme étant des zones comportant un certain nombre de caractéristiques environnementales communes. Ces caractéristiques sont stockées dans des couches d'informations de SIG. La sélection des couches nécessaires pour obtenir des unités écologiques acceptables (sémantiquement et quantitativement) nécessite de mettre en place un processus d'optimisation entrainant des modifications des couches (relaxation). Les unités sont ensuite utilisées par les biologistes afin de placer des placettes de surveillances des différents écosystèmes.
- ✓ Production : modélisation et formalisation du problème, définition de l'algorithme de résolution, fusion et sélection de l'information pertinente.

Projets de Master recherche 2ème année (ou assimilé)

Frank Duhamel (septembre à décembre 2007)

Titre : *Orthorectification et recalage d'images aéroportées de la Guadeloupe de 1955*

- ✓ Projet long
- ✓ Encadrement : 100 %
- ✓ Détails : l'objectif du projet était de résoudre un problème technique d'orthorectification et de recalage d'images afin de mettre en correspondance des images papier de la Guadeloupe datant de 1955, et fournies par le Parc National, avec des images datant de 2004. Cette mise

en correspondance devait servir à répertorier les modifications du littoral de l'archipel ainsi que les extensions ou régressions de la mangrove. Les traitements devaient être effectués au sein d'un SIG. Les images acquises par des satellites ou des avions doivent être projetées dans un système commun au SIG afin de permettre la correction des erreurs de prise de vue. Pour réaliser l'étude, deux problèmes sont à résoudre : le recalage des images de 1955 entre elles et la projection des images dans le système de projection de 2004.

- ✓ Production : mosaïque des photos aériennes.

Gary Poinin et Hulrick Kodaday (octobre-novembre 2010)

Titre : *Définition de l'ontologie associée aux couverts végétaux de la Guadeloupe*

- ✓ Projet long
- ✓ Encadrement : 100 %
- ✓ Détails : le projet réalisé consistait à fusionner des ontologies existantes dans le domaine public avec des ontologies définies par des chercheurs du laboratoire DYNECAR (*Alain Rousteau*) afin de produire une ontologie unique utilisable à la fois pour la classification supervisée des couverts forestiers et pour l'optimisation de la sélection d'information (thèse de *Wilfried Segretier*).
- ✓ Production : ontologie hiérarchique sur les couverts forestiers.

Stages de Master recherche 1ère année

Frank Duhamel et Manuela Minatchy (avril 2007 à juin 2007)

Titre : *Fusion d'images supervisée par paquets d'ondelettes*

- ✓ Situation actuelle : Manuela Minatchy est en poste à la cellule SIG de la région Guadeloupe.
- ✓ Encadrement : 50 %
- ✓ Détails : un des problèmes identifié et partiellement résolu dans le cadre de la thèse de Mohamed Abadi concernait l'erreur géométrique et radiométrique engendrée par les techniques de fusion d'images satellites de résolution spatiale différente. Cette fusion est opérée pour combiner les hautes résolutions spatiales et spectrales des images. Des travaux de comparaison de différentes transformations en ondelettes ont été menés dans le cadre de ces deux stages.
- ✓ Production : développement d'une méthode de fusion

Ralph Vital (avril 2008 à juin 2008)

Titre : *Optimisation de l'espace couleur hybride pour la classification d'images*

- ✓ Encadrement : 50 %
- ✓ Situation actuelle : en thèse de mathématiques à l'UAG
- ✓ Détails : le choix de l'espace couleur de projection d'une image a beaucoup d'importance

pour les étapes de classification et de segmentation. La répartition des données n'est pas la même en fonction des espaces et certains peuvent faciliter les traitements. Parmi les espaces couleurs, les espaces hybrides sont les plus prometteurs car ils sont construits dynamiquement pour chaque image à traiter. L'objectif de ce stage était la mise en place d'un algorithme d'optimisation de l'espace couleur hybride.

- ✓ Production : maquette Matlab de l'algorithme d'optimisation

3.7 Projets et collaborations

- ✓ **2004-2011**. Montage et suivi du projet CESAR (2004-2007) dans le cadre du programme Européen INTERREG III'b Espace Caraïbes et CESAR Volet II (2009-2011) dans le cadre du programme Européen INTERREG IV (Budget pour l'UAG : 109 000€. Budget global : 300 000€). Mon rôle dans ce projet a été : la définition du projet lui même, la recherche de partenaires, la rédaction du dossier de demande de financement, la réalisation du projet, la rédaction des différents rapports, la gestion du budget et des remontées de dépenses. Ce projet a donné lieu à des collaborations avec l'université de Moncton au Canada et l'université de Curitiba au Brésil.
- ✓ **2005-2007**. Signature de conventions avec la DAF (2005), le Conseil Régional de la Martinique (2006) et le Parc National (2007) pour l'échange de données géo-référencées.
- ✓ **2007**. Participation au projet PARAGE associant SpotImage, le CIRAD, l'IRD et SIGBEA. Production de cartographies réduites d'espaces forestiers pour l'évaluation de l'apport de la télédétection dans la gestion des espaces naturels sur la Guadeloupe.
- ✓ **Depuis 2010.** Collaboration avec *Alain Rousteau* du laboratoire DYNECAR pour la mise à jour de la cartographie des espaces forestiers de la Guadeloupe. Etude comparative de plusieurs méthodes de classification. Modélisation des données par des ensembles flous. Fusion d'informations floristiques et topologiques.
- ✓ **Depuis 2010**. Collaboration avec le laboratoire XLIM-SIC, UMR CNRS 6172 de l'Université de Poitiers. La collaboration porte sur l'étude du critère d'information pour la sélection d'attributs. Application à la sélection d'attributs de textures dans un espace de grande dimension (supérieure à 2000 attributs) pour la classification de couverts forestiers.
- ✓ Dépôt d'un dossier en collaboration avec Artur José Gil en février 2011 dans le cadre de l'appel à projet NetBiome. Projet ECOSENSING. Partenariat Espagne, Portugal, France. Budget: 280 000 €.
- ✓ Dépôt d'un dossier en mai 2011 dans le cadre de l'appel à projet Outre Mer. Projet CIEF : Cartographie Interactive des Espaces Forestiers. Partenariat : DYNECAR (UAG), XLI-SIC (Université de Poitiers). Budget : 43 000 €. L'objectif du projet est de mettre à jour la cartographie des espaces forestiers de la Guadeloupe en utilisant des classifications supervisées.
- ✓ Projet de dépôt d'un dossier courant 2012 dans le cadre du programme européen INTERREG IV Espace Caraïbes pour la définition d'un réseau de SIG sur les Antilles.

3.8 Publications

Revues Internationales

[1] Navy, P., and Page, V., and **Grandchamp, E.,** and Desachy, J., *Matching two clusters of points extracted from satellite images,* Pattern Recognition Letters Journal, Special Issue of Pattern Recognition and Remote Sensing, Vol. 27, No 4, pp. 268-274, 2006

Revues Nationales

[2] Abadi, M. et **Grandchamp, E.,** *Classification de couverts végétaux par analyse de textures couleurs d'images satellites haute résolution*, revue Traitement du Signal, Vol. 26 No 2, mars 2009, Numéro spécial Télédétection pour la surveillance et la gestion de l'environnement

[3] **Grandchamp**, **E.,** *Coopération inter-couche dans un algorithme hybride*, Journal Européen des Systèmes Avancés, 2005

Chapitres de livres

[4] **Grandchamp, E.,** *Raster Vector Integration Within GIS*, Chap. XX, The Geographical Information Sciences, (publication décembre 2011), Intech, ISBN 979-9-5330-7419-0.

[5] Abadi, M., and **Grandchamp**, **E.**, *Large deviation spectrum estimation in two dimensions*, Signal Processing for Image Enhancement and Multimedia Processing, Vol. 31, Chap. 28 pp. 323-333, Springer, 2007, ISBN 978-0-3877-2500-0

[6] **Grandchamp**, **E.,** *An hybrid approach to real complex system optimization,* International Federation for Information Processing, Vol. 172, Chap. 10, pp. 169-200, Springer, Kluwer, 2005, ISBN 038-7-2404-89

[7] **Grandchamp, E.**, Mezzadri-Centeno, T., Hervet, E., *Recensement forestier : projet CESAR*, Ecosystèmes forestiers des Caraïbes, Chap. 10, pp 185-195, 2009, ISBN 978-2-8111-0090-2

Conférences internationales

[8] Segretier, W., and Collard, M., and **Grandchamp**, **E.,** *An heuristic-based approach for merging layers information in a GIS*, GeoProcessing 2012

[9] Abadi, M., **Grandchamp**, **E.,** Alata, O., Olivier, C., Khoudeir, M, *Information criteria performance for feature selection,* CISP 2011

[10] **Grandchamp**, **E.**, *Specification for a Shared Conceptual Layer in GIS,* GeoProcessing 2011

[11] **Grandchamp**, **E.**, *Raster-vector cooperation algorithm for GIS*, GeoProcessing 2010

[12] **Grandchamp**, **E.**, *Automatic delineation of forest ecosystems by combining GIS layers and Remote Sensing images*, ICECS 2009

[13] Abadi, M., **Grandchamp, E.,** and Khoudeir, M., *Improving spatial and spectral resolution of satellite images*, VIPIMAGE 2009

[14] **Grandchamp**, **E.**, *GIS information layer selection directed by remote sensing for ecological unit delineation*, IGARSS 2009

[15] **Grandchamp, E.,** and M. Abadi, *Hybrid color space choice: an optimization review for cost/efficiency trade-off*, SITIS 2009

[16] Abadi, M., and **Grandchamp**, **E.**, *Colour space influence for vegetation image, SPIE*, Cardiff 15-18 Septembre 2008

[17] Abadi, M., and **Grandchamp**, **E.**, *Texture features and segmentation based on multifractal approach*, Lecture Notes in Computer Science 4225, Congress On Pattern Recognition (CIARP), *page 297-305*, 2006

[18] Abadi, M., and **Grandchamp**, **E.**, *Large deviation spectrum estimation in two dimensions*, Signal-ImageTechnology & Internet Based Systems (SITIS) 2006

[19] Abadi, M., and **Grandchamp**, **E.**, *Legendre Spectrum for texture classification*, IEEE Conference on Signal Processing (ICSP) 2006

[20] Abadi, M., and **Grandchamp**, **E.**, *Estimation of the large deviation spectrum,* CAS 2006

[21] **Grandchamp, E.**, *An hybrid approach to real complex system optimization : Application to satellite constellation design,* High Performance Computational Science and Engineering (HPCSE), 2004

[22] **Grandchamp**, **E.**, and Marthon, P., *Driving Segmentation and recognition phases using multiscale characterization*, IGARSS'2003

[23] **Grandchamp, E.,** and Charvillat, V., *Metaheuristics to Design Satellite Constellations*. MIC'2001

[24] **Grandchamp, E.,** and Charvillat, V., *Integrating Orbit Database And Metaheuristics To Design Satellite Constellation*. ICAI'2000

[25] **Grandchamp, E.,** and Charvillat, V., *Satellite Constellations Optimization With Metaheuristics*. Euro'2000

Conférences nationales

[26] **Grandchamp**, **E.**, Regis, S., and Rousteau, A., *Génération de classes de transition vectorielles par recouvrement de classes floues*, RFIA, 2012

[27] Abadi, M., Alata, O., Olivier, C., **Grandchamp, E.**, et Khoudeir M., Critères d'information pour la sélection de variables. Traitement et Analyse de l'Information, Méthodes et Applications (7ème TAIMA), ARTS-PI éd., Vol 2, pp 38-44, Hammamet (Tunisie), Oct. 2011

[28] Abadi, M., **Grandchamp**, **E.**, Alata, O., Olivier, C., Khoudeir, M., *Sélection des variables optimales par optimisation multi-objective de l'information mutuelle,* GRETSI 2011

[29] Navy, P., and Pagé, V., and **Grandchamp, E.,** and Desachy, *J., Appariement de deux clusters de points extraits d'images satellite*, ORASIS, 2005

[30] **Grandchamp, E.,** and Marthon, P., *Segmentation et détection d'objets par caractérisation multi-échelle*, RFIA'2004

[31] **Grandchamp**, **E.**, and Hervet, E., and Mezzadri-Centeno, T., *Recensement forestier : présentation du projet CESAR, Colloque International et Pluridisciplinaire*, Les écosystèmes forestiers des Caraïbes, 2005

[32] Abadi, M., **Grandchamp**, **E.,** *Analyse d'images forestières à l'aide du formalisme (multi)fractal*, Colloque International et Pluridisciplinaire, Les écosystèmes forestiers des Caraïbes, 2005

[33] **Grandchamp**, **E.**, and Pagé, V., and Remi, C., and Enée, G., *Analyse de productions graphiques d'enfants*, EIAH 2005

Brevet

[34] Co-dépôt (30%) du brevet ANALYEAST dans le cadre de la thèse de Laurent Manyri et porté par le LAAS UPR 8001 Toulouse 2006

Mémoires

[35] **Grandchamp, E.,** *Quelque contribution à l'optimisation de constellations de satellites.* Manuscrit de thèse de doctorat, *INP Toulouse*, 2001

[36] **Grandchamp, E.,** *Segmentation d'image radars par ondelettes*, Mémoire de stage de DEA, *INPT-ENSEEIHT*, 1998

Perspectives

[37] Abadi, M., Loumi, S., **Grandchamp, E.,** and Alilat, F. *Colour Image Texture Characterizing: Caribbean Forest Classification by IKONOS Images*, International Journal Of Remote Sensing, **(soumis)**

[38] Girdary, L.D., **Grandchamp**, **E.,** Marrama, L., *A two step housing patterns classification for the study of dengue transmission in Guadeloupe*, International Journal of Geographical Systems **(soumis)**

[39] **Grandchamp, E.,** Abadi, M., Alata, O., *An hybrid method for feature selection based on multiobjective optimization and mutual information*, Pattern Recognition **(soumis)**

[40] Abadi, M., Gil, A. and **Grandchamp, E.,** *Assessment of application of Pan-sharpening Methods on Crop and Forest Identification,* Journal on Photogrammetric Engineering and Remote Sensing (**soumis**).

[41] Segretier, W., **Grandchamp**, **E.**, and Collard, M., *Une approche heuristique pour la fusion de couches d'information*, SAGEO, 2012 **(en cours de soumission)**

[42] **Grandchamp**, **E.,** and Rousteau, A., *Forest classification revision using spatial analyse, fuzzy modelling and supervised classification.* (s**oumission prevue début 2012)**. Journal of System Information Sciences (GeoInformatica).

3.9 Responsabilités et fonctions électives

- ✓ **2003-2007**. Membre élu suppléant de la commission de spécialiste section 27. Examen des dossiers de candidature pour les recrutements de 2004, 2005 et 2006.
- ✓ **2003-2010**. Membre élu du conseil du laboratoire GRIMAAG en charge du suivi des finances. Ma contribution au sein du conseil de laboratoire a été de développer un site web et une interface de gestion des finances permettant la répartition du budget (dotation ministérielle, contrats, BQR, etc.), la saisie et le suivi des dépenses, la production de statistiques sur le budget et le bilan des exercices depuis 2003.

- ✓ **2004-2007**. Responsable de la deuxième année d'IUP I2M.
- ✓ **Depuis 2005**. Correspondant SPECIF pour l'académie de la Guadeloupe.
- ✓ **2006-2008**. Responsable de la mise en place d'un service web pour la gestion des services d'enseignement de l'UAG (développement personnel et encadrement de 8 stagiaires niveau L3). Le site web permet la gestion des services (saisie des enseignements, production de fiches de service, recherche des enseignements non pourvus, ...).
- ✓ **2007-2011**. Coresponsable de la maquette de la filière Génie des Systèmes Energétiques (GSE) dans le projet d'école d'ingénieur porté par l'UAG. Ma responsabilité dans ce dossier concerne : la définition du contenu pédagogique de la filière, la relation avec l'INP de Toulouse pour le montage du partenariat.
- ✓ **Octobre 2008 à Mai 2010**. Directeur adjoint du laboratoire *GRIMAAG*.
- ✓ **Depuis 2010** : membre élu du conseil de laboratoire et de la commission finance du *LAMIA*
- ✓ **2012-2015** : membre suppléant élu au *CNU* en 27ème section.

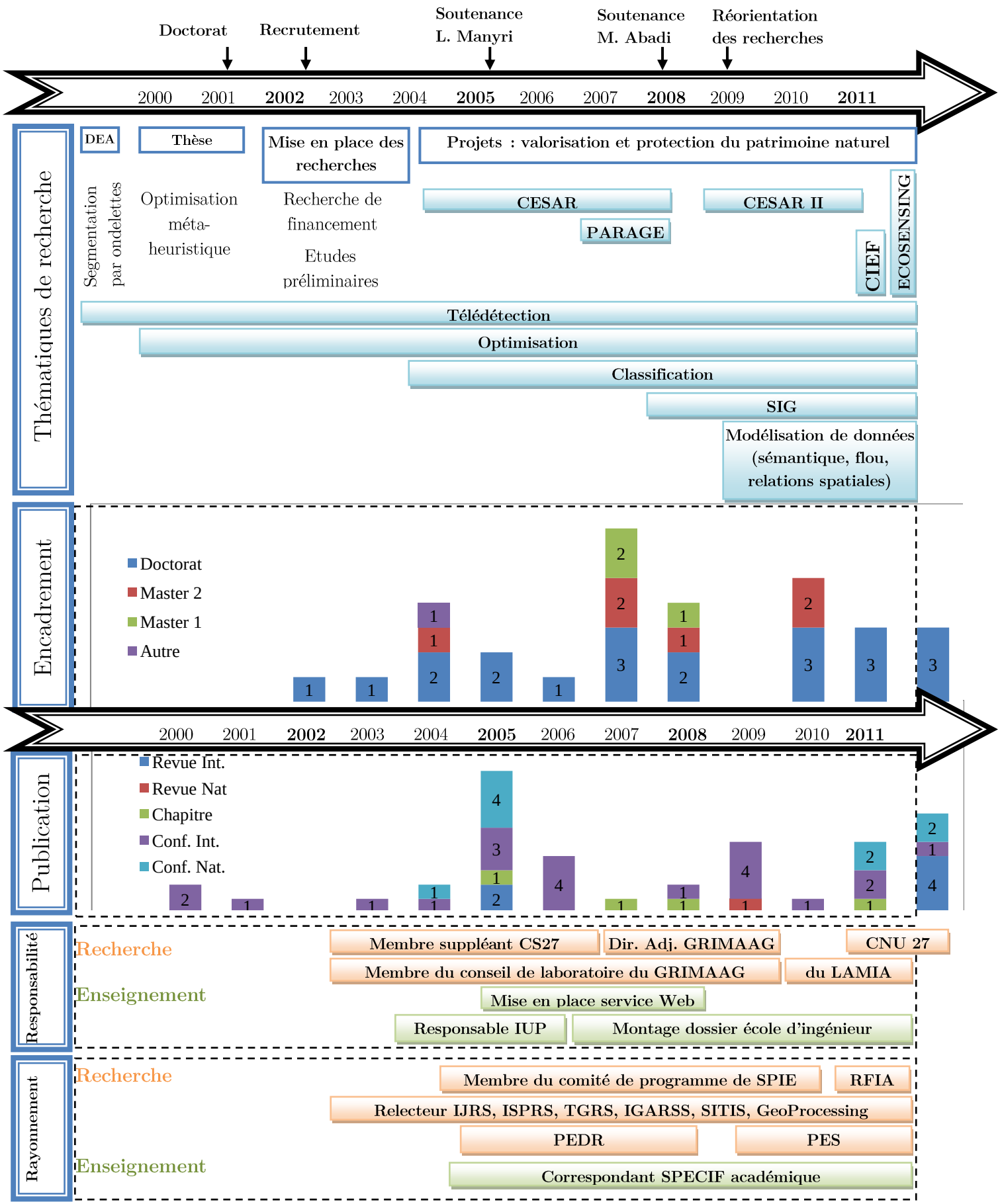
3.10 Animation scientifique

- ✓ **2003-2011.** Relecteur pour les conférences internationales *IEEE International Geoscience And Remote Sensing Symposium* (2003-2011), *International Conference on Signal Image Technology and Internet Based Systems* (2009-2011) et *GeoProcessing* (2010-2011)
- ✓ **2005.** Membre du Jury de thèse de Laurent Manyri, soutenue au LAAS le 14 décembre 2005.
- ✓ **2005-2011.** Relecteur des revues internationales *International Journal of Remote Sensing* (2005-2011), *International Society of Photogrammetry and Remote Sensing* (2007-2008) et *Transaction on Geoscience and Remote Sensing* (2005-2011)
- ✓ **2007-2010.** Membre du comité de programme et relecteur de la conférence *IEEE SPIE Image and Signal Remote Sensing* (RS07)
- ✓ **2008.** Membre du Jury de thèse de Mohamed Abadi, soutenue à l'UAG le 30 juin 2008
- ✓ **2010.** Membre du jury de thèse de Mathias Peroumalnaïk, soutenue à l'UAG le 11 déc. 2010
- ✓ **2011**. Membre du comité de programme de la conférence nationale *RFIA 2012*. Organisation d'un atelier et d'une session en visioconférence entre la Guadeloupe et Lyon.

3.11 Primes de recherche

- ✓ **2004-2008. P**rime d'**E**ncadrement **D**octorale et de **R**echerche (PEDR)
- ✓ **2009-2013**. **P**rime d'**E**xcellence **S**cientifique (PES) (évalué rang A par le ministère en 2009)

3.12 Synthèse chronologique



4 Introduction

Nous signalons que dans l'ensemble du document, les références de [**1**] à [**42**] correspondent aux publications et soumissions personnelles et sont présentées à la page 21, les rapports et mémoires (DEA, Doctorat, etc.) sont présentés en annexe à la page 99.

4.1 Parcours en recherche

Ma carrière de chercheur a commencé en septembre 1997 lorsque je me suis inscrit au *DEA d'Informatique de l'Image et du Langage* (*IIL*) parallèlement à ma troisième année d'école d'ingénieur (*INPT-ENSEEIHT, filière Informatique et Mathématiques Appliquées*). Ce *DEA*, proposé conjointement par l'*Université Paul Sabatier* (*UPS*) et l'*INP* de Toulouse, m'a permis entre autre d'avoir une formation en analyse et traitement d'images, reconnaissance de forme et logique floue. Parallèlement, ma formation d'ingénieur me permettait d'acquérir des compétences en génie logiciel (programmation objet, modélisation) et en mathématiques appliquées (optimisation entre autre).

J'ai donc effectué mes premiers pas dans la recherche durant mon stage de *DEA* et mon stage de fin d'étude entre septembre 1997 et octobre 1998 au sein de l'équipe *image* du *LIMA* de l'*IRIT*. Le sujet que j'avais alors consistait dans un premier temps à développer un algorithme de décomposition d'image sur une base d'ondelettes puis à l'appliquer sur des images radar afin de réaliser un suivi de chaînes de maxima des coefficients d'ondelettes et ainsi caractériser le type de ruptures présentes dans les images afin de localiser les frontières des objets pour segmenter l'image. Ces travaux, réalisés en collaboration avec le *CNES* et le *CESBIO* et dirigés par le Professeur *Philippe Marthon*, m'ont permis d'aborder d'un point de vue théorique et pratique des outils mathématiques complexes [**36**].

Après avoir obtenu mon *DEA* ainsi que mon diplôme d'Ingénieur en Informatique et Mathématiques Appliquées de l'*ENSEEIHT* en 1998, j'ai commencé une thèse de doctorat sur un tout autre sujet. Dirigé par le Professeur *Joseph Noailles*, spécialiste de l'optimisation classique, et encadré par le Professeur *Vincent Charvillat*, expert sur les méthodes d'optimisation meta-heuristiques, ma thèse portait sur l'*optimisation de constellations de satellites*. J'ai effectué ces recherches d'octobre 1998 à décembre 2001 partagé entre le *LIMA* et l'équipe *prospection et mission* d'*Alcatel Space Industries*. Baigné dans un contexte industriel, j'ai pu expérimenter et comparer différentes techniques d'optimisation combinatoires (recherche tabou, algorithmes génétiques (AG), recuit simulé, descente de gradient, etc.) dans des approches globales ou par voisinage pour appréhender un problème complexe. J'ai ainsi pu proposer une modélisation du problème et un algorithme coopératif à plusieurs niveaux pour le résoudre ([**21**], [**23**], [**24**], [**25**], [**35**]).

J'ai ensuite été recruté en septembre 2002 sur un poste de Maître de Conférences à *l'Université des Antilles et de la Guyane* (*UAG*). J'ai intégré le *GRIMAAG* au sein de l'équipe *image* dirigée par le Professeur *Jacky Desachy* et composée à l'époque de cinq Enseignants-Chercheurs. J'ai alors repris mes travaux entrepris en *DEA* sur la télédétection afin de les poursuivre et de les valoriser ([**22**], [**30**]). Durant cette même période j'ai coencadré la thèse de *Laurent Manyri* soutenue en décembre 2005 au *LAAS* à Toulouse sur la segmentation et la

classification de populations microbiennes pour des ferments alcooliques qui a conduit notamment au dépôt d'un brevet ([**34**], [**148**]).

Parallèlement j'ai monté le projet *CESAR* (Classification d'Espèces Arborescentes) dans le cadre du programme européen *INTERREG IIIb Espace Caraïbes* [**43**] avec comme partenaires l'*Université de Curitiba* au *Brésil* et l'*Université de Moncton* au *Canada*. Ce projet a démarré en septembre 2004 avec la thèse de *Mohamed Abadi*, soutenue en juin 2008 [**54**], qui a porté sur la classification de forêt par analyse d'images satellites à très haute résolution. Ce projet a été l'occasion pour moi de tisser un réseau de partenaires avec la signature de conventions au niveau local (laboratoire *DYNECAR*, *Parc National*, *ONF*, *DAF*, etc.) et international (université de Curitiba, université de Moncton) autour d'une problématique particulièrement importante aux Antilles de par l'insularité de ses territoires : la connaissance, la préservation et la valorisation des ressources forestières naturelles. Ce projet a permis de nombreuses publications scientifiques [**2**], [**5**], [**16**] à [**20**], [**31**], [**32**] ainsi que des collaborations avec d'autres partenaires comme *SpotImage*, l'*IRD* et le *CIRAD* autour du projet *PARAGE* [**45**]. Plusieurs stages de master recherche ont par ailleurs été réalisés pour résoudre des problèmes ponctuels liés à cette problématique (caractérisation de textures, fusion d'images, optimisation d'espace couleur, etc.) et une extension du projet est en cours (2009-2011) (projet *CESAR Volet II* [**44**]).

C'est en 2008, après l'aboutissement d'un premier volet important de mes recherches que j'ai ressenti la nécessité de fédérer mes travaux et de trouver un outil d'intégration me permettant de mieux les valoriser. Le choix des *Systèmes d'Information Géographiques* (*SIG*) semblait tout indiqué pour plusieurs raisons : *(i)* les *SIG* sont à la croisée de plusieurs disciplines dont la télédétection, la classification, la visualisation de données spatiales, les statistiques, l'analyse spatiale *(ii)* j'avais eu l'occasion d'utiliser et de mesurer la puissance des *SIG* dans le cadre d'un stage de master recherche pour résoudre un problème de fusion d'information posé par le Parc National de la Guadeloupe *(iii)* l'intérêt régional pour les *SIG*, avec notamment la création d'une cellule SIG au sein du conseil régional en 2007 *(iv)* la nécessité de fédérer leur utilisation au niveau des différents partenaires avec l'organisation en 2007 et 2008 de plusieurs séminaires autour des *SIG*.

J'ai donc consacré en 2008 et 2009 une partie de mes activités à l'intégration des outils et résultats que j'avais obtenus au sein des *SIG* ([**11**], [**12**], [**14**]). Cette étape a fait apparaitre un certain nombre de verrous technologiques des *SIG* qui ont été reconfirmés en juin 2010 lors d'une formation sur les *SIG* que j'ai suivi ainsi que dans une rétrospective sur les *SIG* publiée en 2010 et évoquant les défis à relever dans les 10 prochaines années [**108**]. En effet, l'intégration des outils et méthodes dans une chaîne de traitement permettant de produire une information exploitable (représentation objet des données avec des informations contextuelles et sémantiques) à partir de données brutes image a fait apparaître la nécessité de créer de nouveaux modèles vectoriels pour les données diffuses et d'une intégration plus généralisée de la sémantique de l'information pour la mise en correspondance de l'information ([**4**], [**10**]).

C'est autour de cet axe que j'ai décidé de construire mon programme de recherche pour les années à venir à savoir : la modélisation de données spatiales diffuses dans leur contexte spatial et sémantique.

4.2 Synthèse des contributions

Au travers des différentes thématiques abordées durant mon parcours un certain nombre de contributions ont été faites dans les différentes communautés scientifiques concernées. Cette section présente une synthèse chronologique des principales contributions depuis mon recrutement en tant que Maître de Conférences en septembre 2002.

Les premières contributions présentées ont été réalisées durant le projet *CESAR* et notamment durant la thèse de *Mohamed Abadi*. La première contribution concerne le **choix d'un espace couleur** pour la représentation des images. Ce choix est nécessaire et important car il va influencer certains post-traitements comme la classification [**16**]. L'utilisation d'espaces couleurs classiques a souvent été critiquée dans la littérature et beaucoup d'études ont été menées pour trouver des espaces couleurs hybrides plus discriminant ([**142**], [**201**]). La principale difficulté est de trouver une méthode de recherche rapide et fournissant des espaces hybrides de bonne qualité dans un espace à forte combinatoire. La méthode de recherche proposée dans ce cadre est basée sur une **approche multi-objectif** [**15**] permettant d'obtenir de meilleurs résultats que la méthode de référence [**201**] avec des fondements mathématiques plus rigoureux concernant la convergence.

La seconde contribution concerne la **fusion d'images** satellites de résolutions spatiales et spectrales différentes [**206**]. Notre contribution dans ce domaine est une **méthode hybride** permettant de généraliser et harmoniser les différentes techniques utilisant les concepts de base d'une des catégories de méthodes de fusion (catégorie projection-substitution [**197**]). La méthode proposée est plus stable en termes de qualité de fusion et permet une meilleure conservation de la dynamique des couleurs ([**2**], [**13**] et [**54**]).

Une fois le choix de l'espace couleur hybride et la fusion des images réalisés la problématique devient l'extraction d'informations pertinentes. Dans ce domaine, la première contribution a consisté à explorer un maximum de types d'attributs permettant de décrire l'information fréquentielle, spectrale, structurelle, fractale et multi-fractale ([**17**], [**54**]) et de sélectionner les plus pertinents pour l'application visée avec des approches frontières ([**22**] et [**30**]) et région ([**2**], [**17**], [**18**], [**19**] et [**20**]). La deuxième contribution, plus récente, concerne la **sélection d'attributs** dans un contexte plus général. Elle n'est donc pas focalisée sur l'application en classification de forêts et est applicable dans un cadre général de sélection d'attributs (validation sur les bases de référence UCI [**63**]). La méthode proposée, basée sur une optimisation multi-objectif de **l'information mutuelle** (*IM*) ou d'un **critère d'information** (*IC*), permet d'obtenir des sous ensembles d'attributs plus stables vis-à-vis d'un ensemble de classifieurs que l'ensemble complet en maximisant le taux moyen de bonne classification et en minimisant sa variance. La méthode de sélection alterne deux phases (l'une dite de type *filter* et l'autre de type *wrapper)* permettant de tirer profit des avantages des deux approches et fournit de meilleurs résultats ([**39**], [**9**] et [**28**]) que la méthode de référence ([**171**]) dans le domaine.

A partir des descripteurs précédents, il est maintenant possible de réaliser un apprentissage sur des données terrain ainsi qu'une **classification** des différents espaces forestiers. Des résultats satisfaisants sur la classification des images ont été obtenus en 2008 sur la base des descripteurs sélectionnés ([**2**], [**54**]) mais leur application comporte certaines limites lors de leur utilisation sur de vastes scènes comportant un grand nombre de classes caractérisées par des

descripteurs différents. Les résultats les plus convaincants ont été obtenus dans des contextes particuliers tels que la séparation forêt/Agriculture (projet *PARAGE* [**95**], [**173**]) ou la séparation forêt/forêt dans le cas de frontières bien localisées ([**2**], [**54**]). L'analyse de ces limites a fait émerger la nécessité d'intégrer une information complémentaire pour guider la classification. Dans ce cadre, les principales contributions ont été *(i)* la **constitution d'un dictionnaire** regroupant la description sémantique et numérique des différents types de textures de forêts présents sur la Guadeloupe afin de constituer une base pour l'apprentissage et la classification des images. Les valeurs des descripteurs sont calculées sur différents types d'images afin de proposer une collection la plus complète possible *(ii)* **l'intégration de couches d'information vectorielles** afin de réaliser la classification des couverts forestiers sur l'ensemble de la Basse-Terre (l'une des deux îles composant la partie principale de la Guadeloupe et comportant les principales formations forestières) ([**4**]). Les techniques utilisées dans cette phase de classification (fusion d'information, apprentissage, arbres de décision, etc.) sont généralisables à d'autres contextes et ont été appliqués dans le cadre de la classification d'habitats pour l'identification des facteurs de transmission du virus de la Dengue ([**38**]).

Les différents travaux réalisés sur la classification des forêts (approche raster ou vecteur) ont fait apparaître une problématique de **modélisation de données** liée à la nature même des forêts et généralisable à tout type de **phénomènes transitoires**. Le passage d'un type de forêt à l'autre est un phénomène transitoire dont le gradient dépend de conditions environnementales locales. Ce type de données n'a pas été modélisé et implémenté de manière satisfaisante dans les *SIG* et notamment dans leur représentation vectorielle ([**134**], [**185**]). La première contribution réalisée dans ce cadre concerne la définition d'un **modèle vectoriel flou** adapté aux données ayant des frontières diffuses et permettant une meilleure fiabilité que les modèles vectoriels existants et une plus grande souplesse que les modèles raster ([**4**]). La seconde contribution concerne le passage de modèles multiples flous à un modèle strict unique faisant apparaitre des **classes de transition** identifiées et localisées au lieu de fixer arbitrairement une frontière n'ayant pas de réalité physique ([**37**]). Enfin, la dernière contribution dans le domaine de la modélisation est la définition d'une **couche sémantique** partagée par différentes couches d'information afin de résoudre les problèmes de mise en correspondance d'objets (liés à des manques de précision dans la localisation ou à des représentations différentes d'un même objet) ([**10**]).

Le troisième volet des contributions concerne l'**exploitation des données**. En considérant un certain nombre de couches d'information vectorielles provenant de sources hétérogènes (extraction d'informations provenant d'images, relevés terrains, modélisation de phénomènes à l'aide des modèles flous, etc.) il est possible de produire une information pertinente par croisement de ces différentes couches mais les croisements conduisent bien souvent à une information trop morcelée pour être exploitable. Une simplification de l'information est donc nécessaire et un compromis entre leurs caractères exploitable et représentatif doit être fait. Dans ce cadre, la principale contribution est la définition d'une approche heuristique permettant **d'optimiser l'information sélectionnée** dans une approche multi-objectif faisant intervenir la sémantique de l'information au travers d'une ontologie et sa représentation au travers de sa géométrie et de sa localisation ([**8**], [**41**]). Enfin la dernière contribution concerne le découpage

des données vectorielles unitaires par analyse de l'information raster associée. Cette approche a donné lieu à un **algorithme de coopération raster-vecteur** ([**4**], [**10**] et [**11**]).

Les travaux sur la modélisation des données n'ont été initiés que très récemment (à partir de 2009) et sont toujours en cours notamment au travers de la thèse de *Wilfried Segretier*, d'une collaboration avec *Alain Rousteau* du laboratoire DYNECAR, de travaux sur les méthodes de recouvrement entrepris avec *Sébastien Regis* et sur mes propres recherches. Ils seront décrits plus en détail dans les perspectives de la section 6 car ils constituent véritablement l'axe de recherches privilégié pour les années à venir.

4.3 Organisation du mémoire

Le reste du mémoire s'organise en deux grandes parties : la section 5 présente les principales *contributions* de mon travail de recherche et de mes encadrements organisées selon une chaîne de traitement décrite à la section suivante ; la section 6 présente les *perspectives* à court, moyen et long terme à travers l'axe de recherche que j'ai choisi.

La section 5 se décompose de la manière suivante :

1. Le paragraphe 5.1 présente le contexte dans le quel est définie la chaîne de traitement avec notamment la notion de *Système d'Information Géographiques* (5.1.1), de *raster* et de *vecteur* (5.1.2) ainsi que quelques *verrous technologiques* (5.1.3).
2. nous aborderons dans le paragraphe 5.2 les travaux relatifs à la *préparation des données et à l'extraction de l'information* avant leur intégration dans les *SIG*. Ces travaux regroupent *(i)* des études en *traitement d'images* (5.2.1) avec le *choix de l'espace couleur* le plus adapté pour la représentation d'une image (5.2.1.1) et la *fusion d'images* de résolution spatiale et spectrale différentes (5.2.1.2), *(ii)* des études sur *l'extraction de caractéristiques d'une image* (5.2.2) avec la *caractérisation de ruptures* (5.2.2.1), *l'analyse de textures et de couleurs* (5.2.2.2), (iii) la *sélection d'attributs* (5.2.3), *(iv) des études* sur la *classification d'images satellites* (0).
3. Dans le paragraphe 5.3 nous présentons des travaux relatifs *(i)* à la *constitution d'un dictionnaire* (5.3.1), *(ii)* à la *classification vectorielle* (5.3.2) avec la *mise à jour de la carte écologique* (5.3.2) et la *classification des types d'habitat dans un contexte de dengue* (5.3.2.2), *(iii)* à la *modélisation de l'information* pour la *représentation des structures diffuses* (formations forestières par exemple) dans un *SIG* (5.3.3) *(iv)* et à la *modélisation sémantique d'une scène* (5.3.4) afin de résoudre les conflits de représentation.
4. Enfin dans le paragraphe 5.4 nous détaillons les travaux relatifs à l'*exploitation de l'information* dans un *SIG* avec notamment l'o*ptimisation de la sélection d'information* (5.4.1) et la *coopération raster-vecteur* (5.4.2).
5. Le paragraphe 5.5 présente un schéma synthétique de l'organisation de ces différents travaux.

La section 6 se décompose de la manière suivante :

1. La section 6.1 définit le contexte de cohésion autour des *SIG*.
2. la section 6.2 présente les *perspectives à court et moyen terme* associées à l'exploitation de l'information.

3. La section 6.3 présente les *perspectives à long terme* concernant la *modélisation de l'information* dans les traitements. Ces perspectives constituent l'axe principal visé et entamé par mes récentes recherches.

5 Contributions

Les principales contributions détaillées dans cette section (qui s'échelonnent sur la période 2002-2011) sont organisées selon le schéma de la Figure 1, qui présente de manière synthétique et simplifiée une chaîne de traitement en trois grandes étapes. Un schéma complet de la chaîne de traitement est repris en fin de section (Figure 41 page 81).

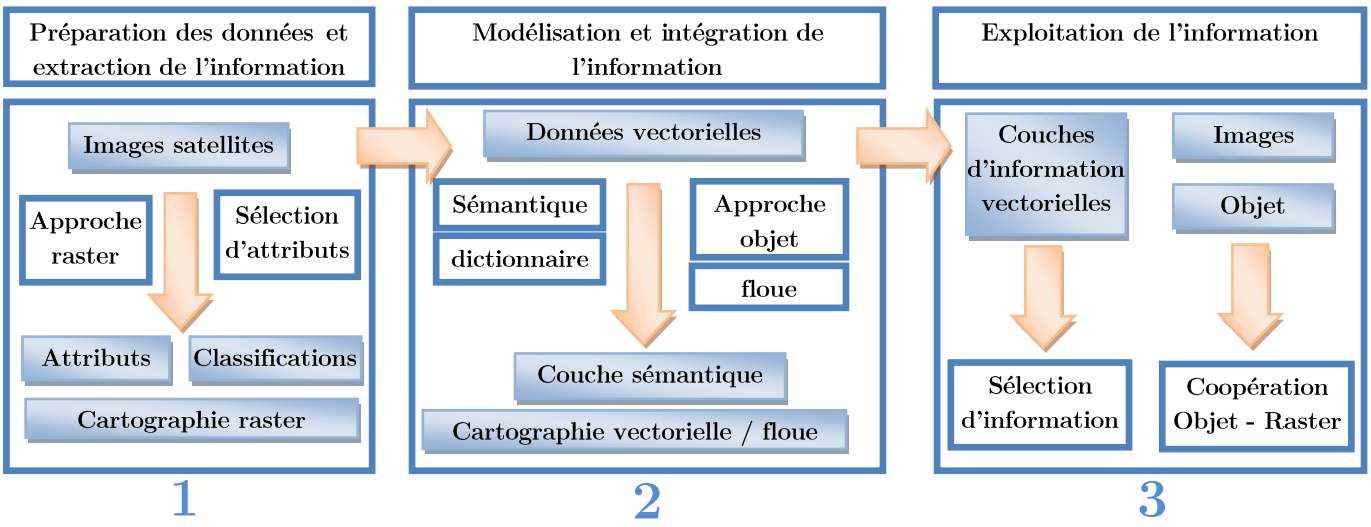


Figure 1 - Schéma synthétique simplifié de la chaîne de traitement

Comme indiqué dans les sections précédentes, cette chaîne de traitement est le fruit de l'intégration depuis 2008 dans un environnement adapté (les Systèmes d'Information Géographiques (*SIG*)) de nombreux outils et de nombreuses approches développés lors de mes recherches. La première partie de la chaîne de traitement permet d'extraire de l'information et de passer des données brutes représentées sous forme d'images (*raster*) à des informations exploitables représentées sous forme vectorielle (*vecteur* ou *objet*). Nous faisons une différence entre le terme *donnée* qui représente des données brutes détachées de toutes connotations sémantiques et/ou contextuelles du terme *information* qui représente une donnée annotée, ayant du sens. Suite aux limitations rencontrées pour représenter l'information extraite précédemment, et également constatées par la communauté *SIG* ([**108**], [**185**]), la seconde partie tente de définir des modèles (dictionnaire, couche sémantiques ou modèles flous) permettant de mieux représenter l'information. Enfin la troisième partie exploite cette information de diverses manières comme la sélection d'information ou la coopération raster-vecteur.

Mais avant d'être détaillée cette chaîne de traitement nécessite d'être replacée dans son contexte (section 5.1) en définissant précisément la notion de *SIG* et les notions de *raster* et de *vecteur*. Nous présentons également quelques verrous technologiques liés aux *SIG* et qui justifient certaines orientations dans les recherches avec notamment l'introduction de la modélisation en

seconde partie de la chaîne de traitement (hétérogénéité des données, mauvaise adaptation des structures de données, etc.). Nous présentons également quelques notions d'optimisation combinatoire (5.1.4) qui seront utilisées de manière transversale dans différentes parties du mémoire.

5.1 Le contexte

Avant d'aborder des problématiques techniques relatives à la chaîne de traitement proposée, nous allons repositionner sa pertinence dans son cadre applicatif. Même si la majeure partie des contributions dépassent le cadre dans lequel elles ont été initialement pensées (sélection d'attributs, modélisation de données diffuses, sélection d'information, etc.) il est important de rappeler le contexte dans lequel les études ont été menées car d'un point de vue financier, technique et humain ces applications ont servi de moteur à l'ensemble des contributions réalisées.

L'application pionnière qui m'a permis de définir à la fois des projets dans le cadre du programme européen INTERREG, de tisser des partenariats locaux, nationaux et internationaux et de faire ressortir des problématiques d'actualité est la classification des forêts des îles de la Caraïbes avec en premier lieu celles de la Guadeloupe. Cette application, a permis d'associer les domaines de l'imagerie, de la classification et de la modélisation qui constituent le triptyque central de mes recherches.

Cette problématique de classification des forêts et son intégration au sein des *SIG* est en adéquation avec un certain nombre de problématiques de la région Caraïbes. En effet, l'exploitation des *SIG* et des données qu'ils contiennent est un enjeu important pour les espaces insulaires sur lesquels on retrouve toutes les problématiques d'aménagement et de surveillance du territoire, de protection et de valorisation des ressources et de développement économique sur un espace relativement restreint qu'il faut gérer de manière efficace. Pour toutes ces raisons, l'ensemble des laboratoires de biologie marine et végétale, physique, géologie, informatique, médecine, etc. utilisent des *SIG* pour effectuer des géo-traitements, des géostatistiques, des simulations de propagation de phénomènes (épidémies de dengue, cyclones, etc.). De même l'ensemble des acteurs de ces régions (conseil régional et général, parc nationaux, *ONF*, *DAF*, *BRGM*) utilisent ces systèmes et rencontrent un certain nombre de problèmes similaires à ceux présentés par la suite.

5.1.1 Les Sciences d'Information Géographiques (SIG)

Le terme *SIG* est une abréviation pour *Système* ou *Science d'Information Géographique* (*GIS : Geographic Information System or Science*). Ce sont donc des Systèmes d'Informations dont les informations stockées et manipulées possèdent une référence spatiale permettant de les localiser à la surface ou à proximité de la terre.

Aujourd'hui les *SIG* ne regroupent plus uniquement des sciences issues de la géographie mais une constellation de domaines dont les problématiques trouvent des déclinaisons au sein des *SIG* ([**105**],[**89**]) (Intelligence Artificielle, traitement et analyse d'images, rendu *2D* et *3D*, télédétection, (géo)mathématiques, (géo)statistiques, logique, théorie des graphs, théorie des ensembles, géométrie, etc.). De même les *SIG* ne sont plus uniquement géographiques mais comportent une dimension thématique, une dimension spatiale (*OD, 1D, 2D* et *3D*) et une

dimension temporelle ainsi que toutes les interactions entre celles-ci. Ceci implique de devoir raisonner et analyser dans les trois dimensions simultanément [**59**] en définissant des métriques dans chacune d'elles ([**100**], [**128**], [**180**]). On parlera ainsi d'ontologie et de sémantique pour la dimension thématique, de distance euclidienne et de géométrie pour la dimension spatiale et de dynamique temporelle pour la dimension temporelle.

Les informations sont regroupées de manière thématique au sein de couches d'information que l'on peut superposer, croiser, ou combiner de différentes manières en respectant des opérateurs matriciels ou ensemblistes.

C'est dans les années 90 que les *SIG* sont passés d'une technologie permettant d'afficher des cartes sur un ordinateur à une discipline à part entière ([**80**], [**212**]). En effet, le simple affichage de cartes thématiques rudimentaires à beaucoup évolué et les outils *SIG* permettent maintenant de réaliser de vraies analyses en combinant information tabulaire et information spatiale : géostatistiques, classification, analyse spatiale (proximité, etc.), modifications géographiques (recouvrement, intersection, etc.), etc.

Néanmoins, un des objectifs finaux de ces *SIG* est de produire une information cartographique exploitable qualitativement (aspect visuel de la carte) et quantitativement (information contenue dans la carte). Ceci a fait émerger des problématiques telles que la visualisation des résultats à distance (cartographie sur le web via des services tels que Google Earth, Geoportail, etc.), l'importance des métadonnées descriptives, la découverte d'informations géographique (*Geographical Knowledge Discovery – GKD* [**162**]) qui est un sous domaine du data mining et plus précisément de la découverte de connaissance (*Knowledge Discovery from Data – KDD*), la nécessité d'outils d'analyse poussés permettant la combinaison de couches d'informations nombreuses et hétérogènes (par leur nature, polygone, ligne, vecteur, raster ou leur origine, résultat de traitements ou processus manuel, etc.).

L'interprétation statique des données est également en train de laisser la place à des outils de simulation dynamiques de phénomènes physiques et temps réels (inondation, cyclone, embouteillage, suivi de flotte, etc.). Des outils complexes d'analyse et de recherche dans des réseaux par exemple (plus court chemin, tournée minimale, etc.) ont été implémentés dans un certain nombre de *SIG*.

Les *SIG* ont donc surpassé le cadre de la géographie pour envahir notre quotidien au travers d'une multitude d'applications (navigation, localisation, surveillance, pour des problématiques de tourisme, aménagement du territoire, sécurité, environnement, etc.). L'analyse du marché de l'industrie géo-spatiale publiée en 2009 fait état d'une augmentation annuelle comprise entre 10 et 20% [**190**]. Aujourd'hui aucune discipline ne peut prétendre être indépendante des données géoréférencées et les nouvelles fonctionnalités concernent donc à la fois un métier (géographie, transport, énergie, etc.), une technologie (visualisation 2D, 3D, web, portabilité, etc.) et une discipline scientifique (optimisation, télédétection, statistique, physique, etc.).

5.1.2 Différentes représentations d'une même réalité

Raster vs Vecteur

En tant que *Système d'information* (*SI*), les *SIG* proposent des structures pour stocker les données. Si l'information tabulaire respecte les structures classiques des *SI*, les données géographiques nécessitent quant à elles des structures de données particulières qui vont dépendre de la nature des données. Deux grandes catégories se distinguent :

1. les données représentées de manière *raster* ([**83**]) c'est-à-dire sous la forme d'une grille de valeurs. On trouve dans cette catégorie : les images représentées sous forme de pixels, le résultat d'un traitement sur des données comme une grille de labels après une classification d'images, une grille d'attributs représentant par exemple les précipitations en chaque point d'un maillage, etc.
2. les données représentées de manière vectorielle (également appelées représentation *objet* ou *vecteur* [**83**]). Ces données sont représentées à l'aide de modèles géométriques simples tels que les points, de lignes, de polygones et de surfaces.

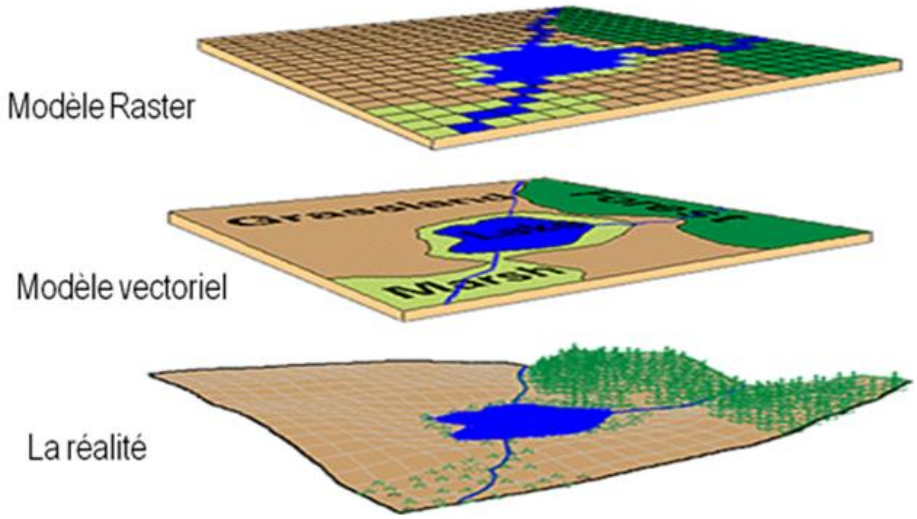


Figure 2 - Raster vs Vecteur (extrait de [117])

Les deux représentations précédentes sont des discrétisations différentes du monde réel (Figure 2) conduisant à des modèles très différents, tant par la nature des structures permettant de le représenter que par les outils permettant de les manipuler. Le débat entre ces deux représentations est né, s'il n'existait pas déjà, avec les *SIG* et a toujours été d'actualité car ces deux catégories possèdent bien évidemment chacune leurs avantages et leurs inconvénients.

Les avantages de la représentation raster sont la simplicité du modèle (grille de valeurs) et la simplicité des algorithmes de combinaison de différentes couches raster (addition, multiplication, etc. de matrices). Le développement des *SIG* orientés raster et des outils d'analyse spatiale ont été soutenus par la grande affluence des images satellites et non par une meilleure adaptation des données raster pour répondre aux différentes problématiques des *SIG*. Les images

satellites constituent en effet une source de données importante (voir trop importante) surtout dans l'analyse spatio-temporelle de phénomènes. Elles permettent de couvrir l'ensemble d'un territoire de manière régulière et ainsi (sous couvert de traitements relativement lourds et complexes) de produire une information dynamique permettant de traiter des problématiques telles que les changements d'occupation du sol ou la mesure d'impact d'un phénomène naturel (inondation, cyclone, tremblement de terre, etc.).

Les principaux inconvénients de la représentation raster sont d'une part liés à leur manque d'information contextuelle et sémantique qui rend ce type de données ignorante par rapport à la nature même des objets qu'elles représentent et d'autre part leur lourdeur d'utilisation liée au grand volume de données qu'elles génèrent. Pour comprendre un peu mieux les limites des modèles raster, on peut voir cette approche comme une représentation du monde sous la forme de regroupements de pixels associés à des vecteurs d'attributs mesurables mais sans leur associer de sémantique ou de contexte. Ces regroupements sont instables spatialement et temporellement puisqu'ils peuvent être faits de manière complètement différente en fonction des critères utilisés, et des images utilisées (date d'acquisition, bande spectrale). Un regroupement peut en effet exister dans une bande spectrale et pas dans une autre ou à un instant et pas à un autre. De la même manière attributs et objets ne sont pas équivalents et un regroupement d'attributs peut ne pas correspondre à un objet physique.

A l'opposé du modèle raster, le modèle vectoriel (dit *objet* ou *vecteur*) offre plus de souplesse dans les traitements et est bien plus facilement rattachable à une information sémantique puisque les objets sont clairement identifiés et localisés ce qui leur donne une persistance spatiale et temporelle. Même si ses propriétés peuvent changer, l'objet reste une entité à part entière rattachée à une sémantique. Les données vectorielles peuvent provenir de sources très différentes telles que le relevé de coordonnées par lecture sur une carte, l'utilisation d'un appareil de mesure type GPS ou encore l'analyse automatique d'autres données (raster ou vecteur). L'emploi de modèles géométriques simples permet de manipuler facilement les objets au travers d'opérations ensemblistes du type union, intersection, etc. Des applications comme la recherche d'un chemin dans un réseau de lignes ou la recherche d'objets par des requêtes spatiales du type *proche-de* sont alors possibles.

Dans l'appréhension d'un problème de caractérisation d'objets, deux approches sont alors possibles :

1. Considérer des regroupements d'attributs communs calculés à partir des images qui formeront des objets.
2. Considérer des objets dont on cherche ensuite à calculer les attributs.

La première conception est celle adoptée dans une approche empirique de découverte du monde. Historiquement un certain nombre de phénomènes ont été détectés, localisés puis identifiés par cette approche (c'est le cas des approches présentées dans la partie extraction d'informations). Mais si on se base sur une connaissance déjà importante de l'environnement dans lequel nous évoluons et que nous tentons de représenter dans un *SIG*, la qualification des objets et phénomènes avant leur quantification par des attributs calculés à partir d'images ou d'autres

sources d'information est une approche plus naturelle et plus efficace (cas de l'exploitation des données et plus particulièrement de la coopération raster-vecteur dans nos approches dans laquelle les objets sont d'abord définis et localisés avant d'être renseignés par une série d'attributs calculés à partir d'un filtrage spatial sur l'image).

5.1.3 Quelques verrous

Suite aux remarques précédentes, nous avons fait le choix d'utiliser les images satellites comme source principale d'information. De par la diversité des images satellites (tant en termes de résolution spatiale que spectrale) l'information qu'elles contiennent est importante quantitativement et qualitativement et les outils issus ou développés dans le cadre de la télédétection vont nous permettre d'en extraire une partie.

Par ailleurs, en axant le débat sur le caractère exploitable et interopérable de l'information la représentation vectorielle se détache très nettement et nous allons tenter de représenter au mieux l'information dans des modèles vectoriels classiques ou spécifiques.

Le passage d'une donnée raster à une information vecteur a été souvent souligné comme étant un axe important de développement des *SIG* ([**65**]).

5.1.3.1 Des données non adaptées aux SIG – Des SIG non adaptés aux données

Les modèles de données proposés dans les *SIG* sont adaptés pour la représentation d'objets bien localisés dans le temps et dans l'espace. C'est le cas de la plus part des objets manufacturés (habitations, routes, véhicules, etc.) ou des concepts définis par l'Homme (frontières, limites administratives, etc.). Mais malheureusement les données géographiques naturelles ou les phénomènes physiques ont très souvent des limites diffuses non représentables à l'aide de figures géométriques simples (points, lignes, surfaces). C'est par exemple le cas pour représenter les limites entre une vallée et une montagne, le lit d'une rivière fluctuant dans le temps, les différents types de sols, les limites d'un cyclone, d'un nuage ou d'une avalanche, ou dans notre cas la limite entre plusieurs formations forestières.

Pour ne pas avoir à garder l'information brute très lourde à manipuler et dénuée d'information contextuelle et sémantique, il faut trouver des modèles adaptés permettant de représenter ces entités avec un maximum de fiabilité.

5.1.3.2 Problème de mise en correspondance des données

Nous nous plaçons ici dans le cas de l'information vectorielle, obtenue directement ou à partir de traitements sur des données *raster*. La plus value des applications utilisant les *SIG* est souvent obtenue par combinaison de différentes sources d'information (population et axes routiers, environnement et pollution, relief et précipitations, etc.). La mise en correspondance entre les entités est basée sur leur localisation. Lorsqu'une même entité est modélisée par des approches différentes ou dans des contextes différents (par exemple une approche *raster* et une approche *vecteur* ou encore deux approches *raster* sur des images différentes ou avec des techniques différentes) on obtient des représentations différentes de la réalité. Ces représentations diffèrent par le positionnement des sommets permettant de délimiter les entités.

L'explosion des applications et des métiers utilisant les *SIG* a engendré une explosion des données avec une hétérogénéité très importante de celles-ci du fait du manque de règlementations. L'hétérogénéité est liée au mode d'acquisition et aux objectifs de l'acquisition.

Par exemple, la précision ne sera pas la même pour positionner une route si elle est modélisée par un service de voirie ou par un service cadastral. De plus, le choix même des sommets permettant de délimiter un objet est souvent arbitraire et difficile à faire lorsque les objets délimités ne sont pas des objets manufacturés (aux contours souvent rectilignes). Dans ce cas, on aboutit à des représentations différentes d'une même réalité sans que l'une soit plus juste ou plus précise que l'autre.

Les différentes représentations de la réalité, engendrées par des acquisitions disparates, vont avoir un impact fort sur les traitements. La base commune qui permet la mise en correspondance des différentes couches thématiques étant le géo-référencement toutes les opérations ensemblistes (union, intersection, etc.) vont propager les erreurs de positionnement. Ces erreurs vont engendrer un bruit parfois important sur l'information. Pour résoudre le problème, des techniques permettant de fusionner des sommets ou de supprimer des entités ont été proposées [**94**] mais uniquement en se basant sur une information spatiale (proximité des sommets, taille des entités, etc.) en définissant par exemple des règles pour fixer la topologie générale d'une scène (en éliminant les recouvrements des entités par exemple ou en imposant leur proximité). Mais dans cette approche, la simplification topologique d'une scène n'est possible qu'en perdant en précision sur les sommets modifiés. On ne peut en effet parler que de simplification car les modifications sont apportées à toutes les entités ne respectant pas les règles définies sans tenir compte de leur sémantique. Le modèle de la scène ainsi obtenu n'est donc pas plus fiable (précis) que le précédent si ce n'est du point de vue topologique.

Par ailleurs, la notion même de représentations différentes d'une même entité ou partie d'entité n'est pas prise en compte dans l'approche topologique. Pour apporter des éléments de réponse à ce verrou technologique, nous avons défini la notion de *couche sémantique* représentant les concepts rattachés aux objets et leurs relations spatio-temporelle et sémantique. Cette représentation d'une scène, permet par exemple de choisir la représentation à conserver lors d'un conflit entre deux représentations (en fonction de la précision de l'entité ou de sa sémantique, l'entité *rue* provenant d'une couche thématique sur les voiries sera donc préférée à une entité *rue* provenant d'une couche cadastre).

5.1.4 L'optimisation en toile de fond

Que ce soit dans la recherche d'espaces couleurs hybrides, dans la sélection d'attributs ou encore dans la sélection d'information, nous sommes confrontés à des problèmes d'optimisation dont l'espace de recherche (E) est fortement combinatoire ce qui empêche toute recherche exhaustive.

Pour résoudre le problème, nous faisons appel à plusieurs méthodes de recherche que j'ai eu l'occasion d'expérimenter durant ma thèse [**35**] et qui ont conduit à des publications combinant différentes approches ([**3**], [**6**], [**23**], [**24**], [**25**]). On trouve :

1. des approches locales par voisinage telles que les méthodes séquentielles présentées dans la section 5.2.3. Le principe est de faire évoluer une ou plusieurs solutions en recherchant son successeur dans un voisinage autour de la solution [**194**].
2. des approches globales telles que les *Algorithmes Génétiques* (AG) utilisés dans la section 5.4.1 qui font évoluer une population de solutions [**88**].

Par ailleurs, nous abordons très souvent les problèmes d'optimisation par des approches multi-objectifs [**69**] c'est-à-dire optimisant plusieurs critères à la fois. Si avec une optimisation mono-objectif il est simple de comparer deux solutions entre elles puisque l'on dispose d'une relation d'ordre totale sur E basée sur l'évaluation du critère, avec une optimisation multi-objectif, cette relation d'ordre n'est que partielle et certaines solutions ne peuvent être comparées.

En effet, si on se place dans le cas, non restrictif, de la minimisation d'un ensemble de p critères $\{C_1, \dots, C_p\}$, on définira la relation d'ordre partielle de domination par :

- une solution S_1 domine une solution S_2 selon le critères $i \in [1,p]$ si et seulement si : $C_i(S_1) < C_i(S_2)$ (une solution S_2 est donc non dominée par une solution S_1 selon le critère i si $C_i(S_1) \geq C_i(S_2)$).
- une solution S_1 domine une solution S_2 si et seulement si : $\forall i \in [1,p]\ C_i(S_1) \leq C_i(S_2)$ et $\exists k \in [1,p]\ C_k(S_1) < C_k(S_2)$ (une solution S_2 est donc non dominée par une solution S_1 si $\exists i \in [1,p]\ C_i(S_1) > C_i(S_2)$ ou si $\forall i \in [1,p]\ C_i(S_1) = C_i(S_2)$).
- une solution S_k est non dominée dans E si $\forall S \in E,\ S$ ne domine pas S_k donc si $\forall S \in E,\ (\exists i \in [1,p]\ C_i(S_k) < C_i(S)\ ou\ \forall i \in [1,p]\ C_i(S_k) = C_i(S))$ (ou encore si $\nexists S \in E,\ \forall i \in [1,p]\ C_i(S) \leq C_i(S_k)$ et $\exists k \in [1,p]\ C_k(S) < C_k(S_k)$)

L'ensemble des solutions non dominées est appelé ensemble Pareto optimal. La Figure 3 illustre le principe de domination et de non domination dans le cas de la minimisation de deux critères.

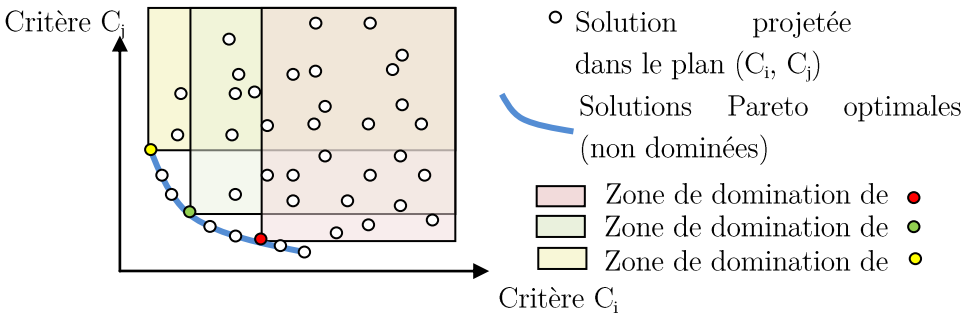


Figure 3 – Ensemble de Pareto pour deux critères

L'avantage de l'optimisation multi-objectif est de pouvoir conserver plusieurs solutions ayant des caractéristiques différentes puisque non dominées selon des critères différents. L'utilisateur peut alors utiliser ses propres critères décisionnels pour sélectionner une solution comme par exemple l'aspect visuel des données dans le cas de la sélection d'information.

5.1.5 Bilans

La structuration de mes contributions et perspectives est donc basée sur la difficulté de gérer plusieurs représentations d'une même réalité :

1. En associant les données raster et vecteur plutôt que de les opposer au travers de l'extraction d'informations pertinentes d'une des plus grandes sources d'information spatiale (les images satellites) tout en développant des modèles de données permettant de les représenter sous le format le plus souple et le plus adapté aux traitements spatiaux (le format vecteur).

2. En tentant de résoudre les problèmes de mise en correspondance (liés aux imprécisions ou aux différences de représentation) existant entre plusieurs représentations d'une même entité par l'introduction d'une couche sémantique.

A partir de cette base, l'exploitation de l'information est facilitée puisque la sélection d'information (parmi une quantité importante de couches d'information vectorielles ou au sein même d'une image) peut maintenant se faire de manière plus efficace en se basant sur la sémantique. Un axe de recherche sur la modélisation des données au sein d'un *SIG* présente donc beaucoup d'intérêt.

5.2 Préparation des données et extraction d'information

Cette partie présente des outils sur la préparation des images (5.2.1 changement d'espace couleur, fusion d'images) avant de calculer des caractéristiques des pixels la composant (5.2.2 analyse fréquentielle, structurelle, fractale, etc.) pour ensuite extraire des informations exploitables comme une labellisation des pixels (5.2.4 application d'un classifieur).

5.2.1 Traitement d'image

5.2.1.1 Choix de l'espace couleur le plus adapté

L'objectif de cette première étape est d'augmenter la discrimination visuelle entre les couverts végétaux et ainsi faciliter l'étape de classification. Le choix de l'espace couleur pour la représentation d'une image n'est ni sans conséquence sur les traitements ultérieurs ni facile à réaliser. La Figure 4 présente la projection de 6 classes dans 4 espaces couleur différents. On peut remarquer les différences de représentation et la forte corrélation des composantes de l'espace RVB (Figure 4 d).

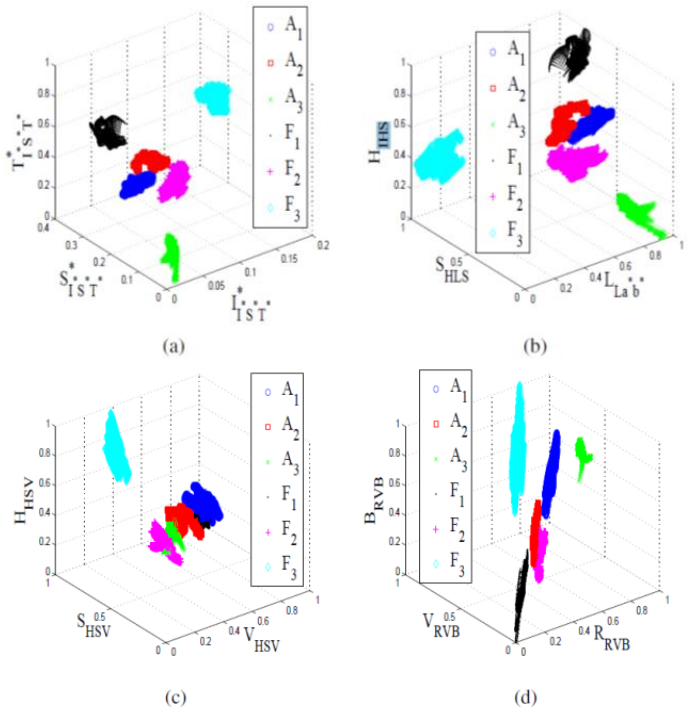


Figure 4 – Projection de différentes classes dans différents espaces couleurs (a) IST (b) LAB (c) HSV (d) RVB

Il existe une multitude d'espaces classiques (20 standards [**74**], [**131**], [**133**], [**141**], [**198**]) dont la plupart ont été démontrés comme peu adaptés pour certains applications comme la classification et la segmentation ([**111**]). De ce fait, on a vu l'apparition d'espaces couleur hybrides ([**142**], [**200**], [**201**]), construits à partir d'un nombre arbitraire de composantes prises parmi les composantes des espaces classiques. Mais le choix de l'espace couleur le mieux adapté pour une application et une image donnée nécessite de définir à la fois des critères de sélection et une méthode de recherche (la forte combinatoire rendant impossible une exploration exhaustive : le nombre total de combinaisons possibles d'espaces construits à partir de N composantes est 2^N avec $N > 30$).

De manière classique ([**77**], [**106**]) la maximisation du *Pouvoir Discriminant* (*PD*) (différentes expressions existent faisant intervenir une minimisation de la *variance intra classe* et une maximisation de la *variance inter classe*) et la minimisation de la *Corrélation* (*Corr*) des composantes couleurs sont les deux principaux critères utilisés.

De même la méthode de recherche classique de l'espace hybride est une approche itérative consistant à alterner l'optimisation des deux précédents critères [**201**]. Cette approche n'a pas de fondements mathématiques rigoureux puisqu'elle fait l'hypothèse implicite et fausse que les deux critères sont indépendants et n'a aucune garantie de convergence vers l'optimum global (approche locale monocritère). Nous illustrons dans [**15**] l'échec de convergence de cette méthode sur un exemple simple et proposons de résoudre le problème par deux approches locales basées sur une optimisation multi-objectif des deux critères : un algorithme glouton (A_1) et un algorithme avec retour (A_2). Nous approchons ainsi les solutions Pareto optimales vis-à-vis du pouvoir discriminant et de la corrélation. Les deux algorithmes sont détaillés dans [**15**] et [**54**].

L'avantage de l'optimisation multi-objectif est de traiter les deux critères simultanément plutôt qu'en les alternant et de renvoyer un ensemble de solutions non dominées plutôt qu'une solution unique.

Pour valider ces approches et s'assurer que la recherche locale converge vers des solutions optimales ou sous optimales, une comparaison des pseudo surfaces de Pareto trouvées avec les surfaces de Pareto calculées à partir de tous les espaces hybrides possibles pour un nombre restreint de composantes a été réalisée sur des exemples test. Comme illustré dans la Figure 5, l'approximation des ensembles de Pareto, obtenue par les deux algorithmes précédents (A_1 et A_2) est de très bonne qualité puisqu'ils sont quasiment confondus avec l'ensemble de Pareto. On remarque également que les espaces couleurs standards (*RGB* et *IHS*) ne sont pas Pareto optimaux. Il en est de même pour l'espace couleur hybride renvoyé par la méthode de [**201**] (étoile verte sur la Figure 5) qui est dominée par la solution entourée d'un cercle discontinu (pouvoir discriminant légèrement plus élevé et corrélation deux fois moins grande).

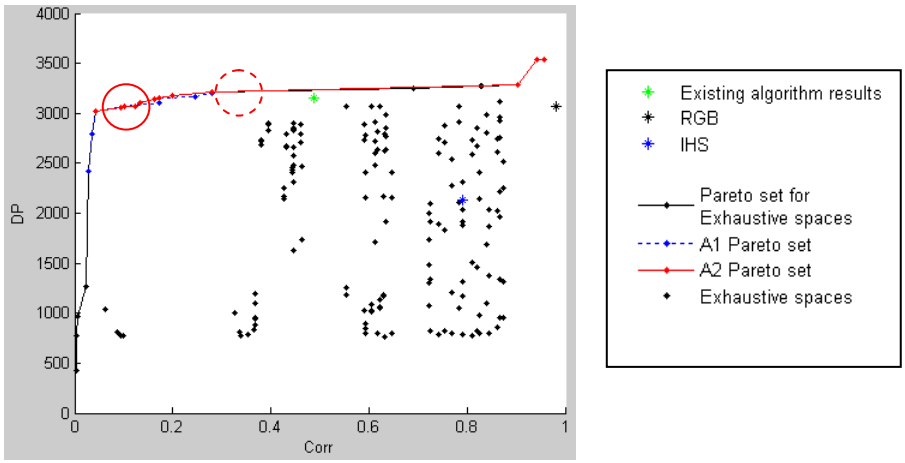


Figure 5 – Comparaison des algorithmes de recherche d'un espace couleur hybride

Ces résultats tendent à montrer une certaine stabilité des surfaces de Pareto à partir d'un certain nombre de composantes dans le sens où les solutions Pareto optimales à $n+1$ composantes semblent majoritairement construites à partir des solutions Pareto optimales à n composantes. Mais ces résultats seraient à vérifier pour des espaces de plus grande dimension pour lesquels la recherche exhaustive est particulièrement longue à calculer.

Réduction de l'espace de recherche

D'un point de vue purement calculatoire, la taille et les composantes d'un espace couleur hybride ont peu d'importance du moment qu'il possède un pouvoir discriminant élevé et/ou une corrélation faible. Mais dans certaines applications, comme la visualisation de données ou certains post-traitements (la fusion d'images dans notre cas) les espaces couleurs retournés devront respecter certains critères (nécessité de trois composantes interprétables une visualisation par exemple). Notre recherche va donc se focaliser sur certains espaces couleurs hybrides et notamment ceux accessibles à partir des différentes expressions des composantes couleurs *Intensité*, *Teinte* et *Saturation (IST)* ([**74**], [**99**], [**107**], [**131**], [**198**]). Ces espaces couleurs non linéaires possèdent de bonnes propriétés car leurs composantes ont une signification physique permettant de séparer l'information : la composante I permet par exemple de calculer des attributs sur les textures de l'image et les composantes S *et* T sur la couleur. La Figure 6 présente l'ensemble des espaces hybrides ayant entre 1 et 3 composantes projetés dans le plan (*Corrélation, Pouvoir Discriminant*). On constate que les espaces de type (*Intensité, Saturation, Teinte*) ont une surface de Pareto de très bonne qualité (ligne bleue), puisque parmi l'ensemble des espaces hybrides seuls quelques espaces sont non dominés par cette surface et certains espaces *IST* sont sur la courbe de Pareto globale. On constate également que les espaces couleur classiques sont loin de la surface de Pareto (HSV et RVB par exemple dans la Figure 6) et l'algorithme de référence (ligne verte) converge vers une solution dominée (le couple (0.5108, 68.10) domine le couple (0.5762, 64.05)) d'une part et d'autre part la solution retournée possède 4 composantes alors que les espaces de type *IST* n'en possèdent que 3.

Etant donnée la forte réduction de l'espace de recherche nous pouvons envisager une recherche exhaustive et une production de la surface de Pareto complète pour les espaces de type

IST ce qui permet d'obtenir des solutions de très bonne qualité pour un temps de calcul acceptable. En effet, il y a 252 espaces à explorer, comparativement aux 87 espaces parcourus dans le cas de [**201**] mais sans garantie de convergence vers un optimum de Pareto.

Les références présentant ces travaux sont [**15**], [**16**] et [**54**]. Ils ont été réalisés durant la thèse de *Mohamed Abadi* et le stage de Master 1ère année de *Ralph Vital*.

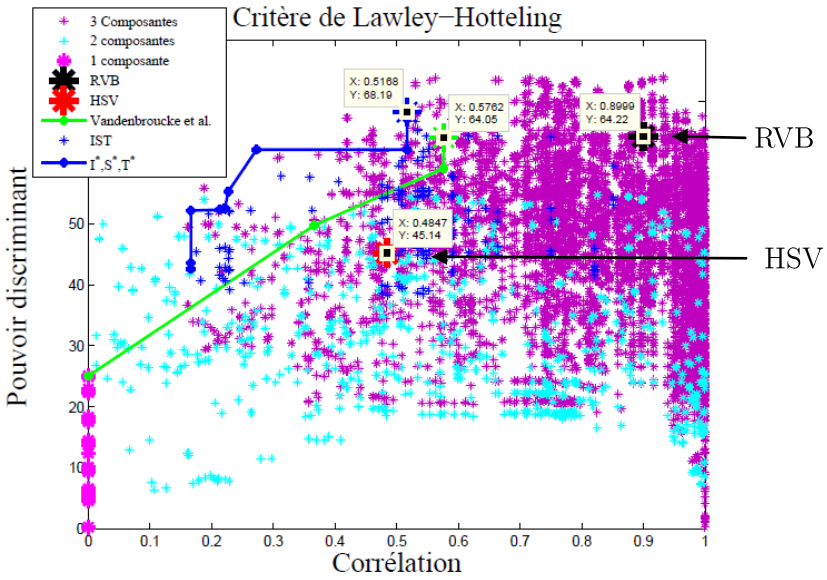


Figure 6 – Comparaison des espaces hybrides

5.2.1.2 Fusion d'images

L'objectif de cette deuxième étape est de combiner la richesse spatiale d'une image et la richesse spectrale d'une autre. Les satellites acquièrent simultanément des images à différentes résolutions spatiales et spectrales. Par exemple : le satellite Spot5 acquiert une image panchromatique (1 bande spectrale), de résolution 5m et une image multi-spectrale (rouge, vert, proche Infrarouge, moyen infra rouge) à 10m de résolution ; le satellite *IKONOS* acquiert une image panchromatique de résolution 1m et une image multi-spectrale (rouge, vert, bleu) à 4m de résolution ; Le satellite *QuickBird* acquière une image panchromatique de résolution 0,64m et une image multi-spectrale (rouge, vert, bleu) de 2,4m de résolution.

L'association de l'information panchromatique à haute résolution spatiale PAN_h à l'information multi-spectrale à plus faible résolution spatiale MS_b permet une analyse plus complète et plus fine des scènes observées. Pour cette raison, un processus de fusion d'images est appliqué sur ces deux types d'image afin d'obtenir une image à haute résolution spatiale et spectrale MS_h.

Il existe dans la littérature quatre grandes catégories de méthodes de fusion [**206**]: la projection-substitution, la contribution spectrale relative (méthode de *Brovey* [**138**]), le concept *ARSIS* (modèle 1 utilisé ici [**197**]), les méthodes hybrides. Notre contribution dans ce domaine est une méthode hybride permettant de généraliser et harmoniser les différentes techniques utilisant les concepts de base de la catégorie projection-substitution.

Les techniques de cette catégorie peuvent être classées en deux groupes : *(i)* les Méthodes Perceptuelles (*MP*) qui sont basées sur des changements d'espaces couleurs (passage de *RVB* vers *IHS*, etc.) par des transformations non linéaires ([**106**], [**188**], [**189**], [**196**], [**206**]) *(ii)* les

Méthodes d'Axes Indépendants (*MAI*) basées sur différentes méthodes statistiques (Analyse en Composantes Principales, Transformée de *Karhunen-Loeve*, Transformée en Cosinus Discrète, etc.) fournissant les composantes les moins corrélées possibles [**125**]. Ces techniques sont basées sur l'hypothèse, souvent vérifiée dans l'espace d'acquisition, qu'il existe des corrélations entre les composantes d'une image.

Une des difficultés des méthodes perceptuelles est de trouver les composantes couleurs adéquates qui permettent d'isoler le contenu spectral du contenu spatial de l'image MS_b afin de minimiser la distorsion radiométrique dans l'image fusionnée MS_h. Ainsi depuis l'introduction de ces méthodes, basées sur les trois composantes *Intensité (I), Saturation (S)* et *Teinte (T)*, plusieurs travaux ont été réalisés dans le but d'améliorer la qualité de l'image fusionnée ([**188**], [**189**], [**197**]).

Nous proposons dans ce cadre une approche originale permettant de prendre en considération les spécificités et les particularités des images fusionnées et ainsi de généraliser les méthodes existantes. Nous utilisons pour cela l'espace hybride *I*S*T** renvoyé par la méthode de sélection d'un espace couleur hybride avec notamment la composante I* contenant le plus d'information et donc la plus proche de l'image PAN_h. Cette méthode est appelée Transformation Hybride (*TH)* et est inversible, condition nécessaire pour le processus de fusion par les méthodes perceptuelles.

Au cours de cette fusion, l'image PAN^S_h est obtenue par une spécification d'histogramme entre l'image PAN_h et la composante I* pour restructurer l'intensité et compenser les détails manquant dans la composante I* par ceux de l'image PAN_h. Ceci permet de réduire les différences dues aux conditions d'acquisition.

Le Tableau 2 présente l'indice de qualité (compris entre 0 et 1) obtenu en appliquant différentes méthodes de fusion dont notre méthode (*TH*), les Méthodes d'Axes Indépendants (*MAI*), les Méthodes Perceptuelles (*MP*), la méthode de Brovey résultats montrent une meilleure stabilité de la qualité de la fusion à partir de notre approche ainsi qu'une meilleure conservation de la dynamique des couleurs vis-à-vis de l'image multi-spectrale à basse résolution (Figure 7).

	Techniques de fusion d'images					**Valeur optimale attendue**
	TH	**MP**	**MAI**	**Brovey**	**Modèle 1**	
GQ	**0.93**	**0.92**	**0.85**	**0.91**	**0.92**	**1**
GQ(R)	**0.93**	**0.94**	**0.81**	**0.93**	**0.91**	**1**
GQ(V)	**0.93**	**0.91**	**0.87**	**0.89**	**0.93**	**1**
GQ(B)	**0.93**	**0.92**	**0.87**	**0.90**	**0.93**	**1**

Tableau 2 - comparaison de l'indice de qualité (GQ) des différentes méthodes de fusion sur les composantes *R*, *V* et *B*.

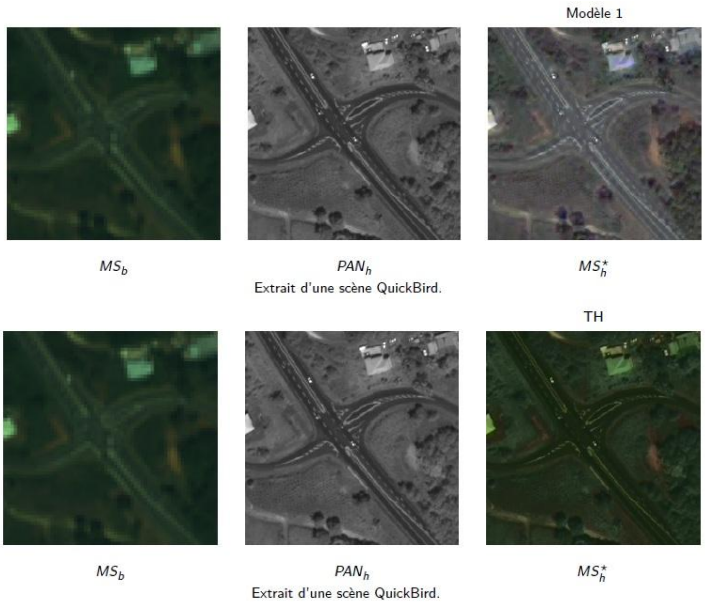


Figure 7 - Comparaison de résultats de fusion d'images entre *TH et le modèle 1*

Les références relatives à ces travaux sont [**2**] et [**54**]. Ils ont été réalisés durant la thèse de *Mohamed Abadi* et les stages de Master 1ère année de *Manuela Minatchy et Franck Duhamel*.

5.2.2 Extraction de caractéristiques

L'objectif de cette troisième étape est de fournir un vecteur de descripteurs calculé à partir de l'image fusionnée précédente permettant de caractériser chaque pixel (ou ensemble de pixels) et ainsi permettre sa labellisation dans une étape ultérieure. Plusieurs types de descripteurs peuvent être calculés et traduisent différentes caractéristiques telles que la couleur, la texture, la forme, etc. Lors de la partition d'une image en zones homogènes, deux approches sont possibles : l'approche région et l'approche frontière. Tandis que l'approche région consiste à caractériser puis localiser des pixels appartenant à un objet, l'approche frontière s'intéresse aux pixels délimitant les objets entre eux. Les deux approches sont duales puisque l'on définit à partir de l'approche région les frontières comme étant l'intersection des objets et à partir de l'approche frontière les objets comme étant localisés à l'intérieur d'une frontière fermée.

Nous présentons ici l'extraction d'attributs en adoptant les deux approches avec, dans un premier temps, une approche frontière permettant la caractérisation des points de rupture dans une image et, dans un second temps, une approche région avec une caractérisation des couleurs et des textures des différents objets recherchés dans l'image. Ces travaux ont respectivement été menés sur les périodes 2002-2004 et 2004-2008.

5.2.2.1 Caractérisation de ruptures

Nous nous plaçons ici dans une approche frontière et cherchons à caractériser puis localiser les différents types de ruptures possibles entre objets afin de guider la segmentation. La caractérisation est obtenue en réalisant un suivi à travers différents niveaux de résolution des chaînes de maxima locaux d'une transformation en ondelettes de l'image.

La transformation en ondelettes est obtenue par projection orthogonale de l'image sur des espaces affines obtenus à partir de la base d'ondelette [**143**]. On parle d'espace d'approximation (*V*) et d'espace de détail (*W*). L'espace d'approximation de niveau *i* contient des signaux plus grossiers que l'espace d'approximation de niveau *i+1*. Plus *i* augmente, plus on s'intéresse aux basse fréquences du signal. L'espace de détail de niveau *i* est la différence d'information entre l'espace d'approximation de niveau *i-1* et l'espace d'approximation de niveau *i* (il sert dont à stocker des hautes fréquences).

Les valeurs des images d'approximation et de détail sont calculées en utilisant un paramètre de dilatation/réduction σ et un paramètre de translation τ. Physiquement, σ permet d'analyser le signal à différentes échelles et τ de parcourir le signal. $W\{f,\psi\}(\sigma,\tau)$ est la valeur du coefficient d'ondelette obtenue par application de l'ondelette ψ au signal *f* avec comme coefficient de dilatation σ et comme coefficient de translation τ.

$$(1) \qquad W\{f,\psi\}(\sigma,\tau)=\langle f(t),\psi_{\sigma,\tau}(t)\rangle \text{ avec } \psi_{\sigma,\tau}(t)=\frac{1}{\sqrt{|\sigma|}}\psi(\frac{t-\tau}{\sigma})$$

où ψ(t) est appelée l'ondelette mère.

Nous utilisons ici des ondelettes bi-orthogonales avec une décomposition discrète dyadique. L'image initiale *I* est donc décomposée en une série d'images d'approximation (V_i) et de détail (W_i^x selon les lignes et W_i^y selon les colonnes).

$$I = V_1 + W_1^x + W_1^y$$
$$(2) \qquad V_i = V_{i+1} + W_{i+1}^x + W_{i+1}^y, 0 \le i \le n$$

La valeur de n dépendant de la taille de l'image.

Les maxima locaux sont ensuite calculés à partir de module des images de détail à chacun des niveaux de résolution (équation (4)).

$$(3) \qquad \left(|W_{i+1}^x|^2 + |W_{i+1}^y|^2\right)^{1/2}, 0 \le i \le n$$

On réalise ensuite un suivi des maxima à travers les différents niveaux de résolution [**145**]. Le comportement des chaînes de maxima (évolution de la valeur des modules, fusion avec d'autres chaînes, disparition ou maintien à travers les niveaux) donne un certain nombre d'indications quant à la nature et à l'importance du point de rupture localisé.

Dans certains cas, pour des points de rupture isolés, on peut déduire de ces chaînes le coefficient de *Lipschitz* (ou exposant de *Hölder*) [**144**] qui permet de caractériser la régularité de la transition. Dans le cas d'images complexes, la proximité des points de rupture et la multitude des types de transition ne permet pas de caractériser la majeure partie des points de rupture par l'expression de ce coefficient. Néanmoins, les chaînes de maxima permettent d'obtenir une signature des transitions et de mesurer leur importance. La signature est liée:

- à la forme générale de la courbe (tableau de variation de la fonction),
- à l'interaction avec les autres chaînes (fusion, disparition, maintien, absorption, etc.),
- aux valeurs prises par le maximum (valeur minimale, maximale, moyenne, etc.),
- aux informations géographiques sur le maximum considéré (position initiale et chemin parcouru sur l'image au travers des niveaux)

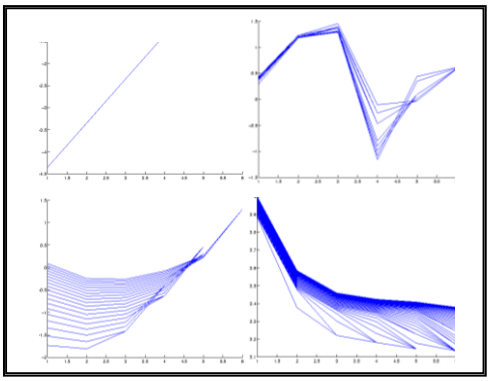


Figure 8 - Exemple de caractérisation

Dans l'exemple de la Figure 8, on peut observer 4 catégories de ruptures caractérisées par des évolutions différentes des chaînes de maxima à travers les différentes échelles. Après caractérisation, on observe la localisation des ruptures correspondantes qui correspondent à des frontières entre objets.

Ces travaux ont donné lieu à deux publications [**22**] et [**30**].

5.2.2.2 Analyse de couleurs et de textures

A travers les travaux présentés dans ce cadre, nous nous intéressons à l'information spectrale (étude de la couleur) et spatiale (étude de la texture).

L'analyse des couleurs d'une image satellite constitue un premier ensemble de descripteurs important. En effet, en calculant les différents moments statistiques (moyenne, variance, entropie, etc.) de la distribution des couleurs de l'image ou d'une fenêtre centrée sur un pixel, on peut caractériser (et donc discriminer et localiser) un certain nombre d'objets ou de phénomènes.

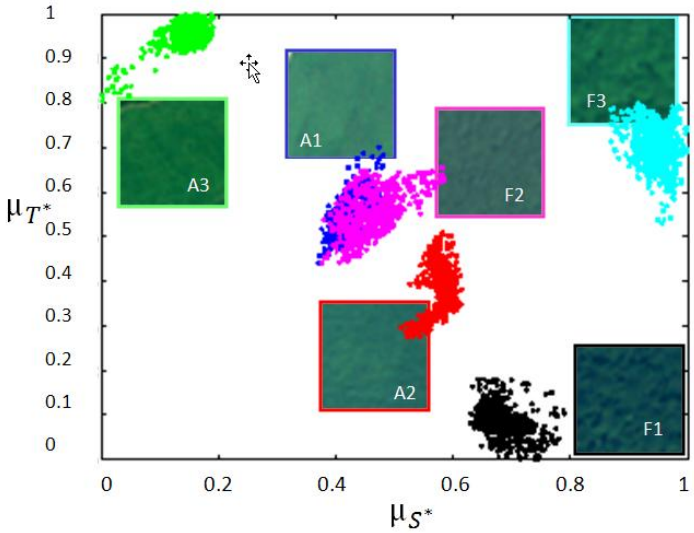


Figure 9 - Extraits d'images représentant des parcelles agricoles (A_i) et des forêts (F_i) et Projection des moyennes des couleurs sur les composantes S et T

La Figure 9 présente trois extraits d'images de parcelles agricoles (A_i) et trois extraits d'images de forêts. Les images sont projetées dans l'espace *IST* (*Intensité, Saturation* et *Teinte*) et la Figure 9 présente une projection de la moyenne des valeurs dans une fenêtre de 8x8 pixels

centrée sur chaque pixel sur les composantes S et T. On constate que dans ce cas de figure, le moment d'ordre 1 (la moyenne) pris sur ces deux composantes permet de séparer les images A_2, A_3, F_1 et F_3 mais pas les images A_1 et F_2.

Les images satellites utilisées pour nos travaux sont des images à très haute résolution spatiale (résolution de 1m pour les images *IKONOS* et 0,6m pour les images *QuickBird*). Ce type d'images fait apparaître des textures au sein même d'objets jusque là visibles dans les images satellites de manière homogène. Ces structures complexes, bien qu'apportant beaucoup d'informations, perturbent les phases de segmentation des scènes en objets et donc la reconnaissance même de ceux-ci.

L'analyse des textures qui composent une image est donc nécessaire pour utiliser cette information pour la segmentation et la classification et ainsi discriminer les objets. Notre application étant la classification des espaces forestiers de la Guadeloupe par analyse d'images satellites et le pouvoir discriminant d'un descripteur dépendant de la composition même de l'image traitée, nous avons étudié les différentes textures de couverts forestiers sur le parc national de la Guadeloupe. Nous avons étudié différentes approches complémentaires afin de caractériser les textures présentes : statistique, fréquentielle, géométrique et fractale.

Concernant les méthodes statistiques nous avons sélectionné 6 descripteurs basés sur les *Matrices de Cooccurrence* tels que définis dans [**54**]. Ces descripteurs ont été sélectionnés car ils sont les plus représentatifs pour nos données.

Les méthodes géométriques permettent de caractériser l'information structurelle et contextuelle des textures, dans ce cadre les *moments de Hu* ont été sélectionnés [**118**] (7 moments calculés et conservés).

Ces descripteurs géométriques sont complétés par des descripteurs issus d'une analyse *fractale* et *multi-fractale* [**118**] plus adaptés au type de textures rencontrées lors de l'analyse des forêts (textures stochastiques). L'idée principale qui nous a conduits à sélectionner ces descripteurs est de pouvoir coupler une description locale de la complexité de la texture avec une information plus globale. Dans ce cadre nous avons sélectionné la *dimension fractale* [**146**] qui permet de traduire la régularité de la structure de la texture, *l'exposant de Hölder* et le *spectre de Legendre* (ou *spectre de singularité* [**116**]) qui sont rattachés à l'analyse multi-fractale.

Enfin les méthodes fréquentielles permettent d'analyser les différentes fréquences des textures. On trouve dans ce registre les méthodes de filtrage telles que les *filtres de Law's* [**130**] (25 descripteurs calculés et conservés) ou *de Gabor* [**147**] (20 descripteurs calculés et conservés).

Pour plus de détails sur l'expression des descripteurs et leurs calculs, le lecteur est renvoyé au manuscrit de la thèse de *Mohamed Abadi* [**54**].

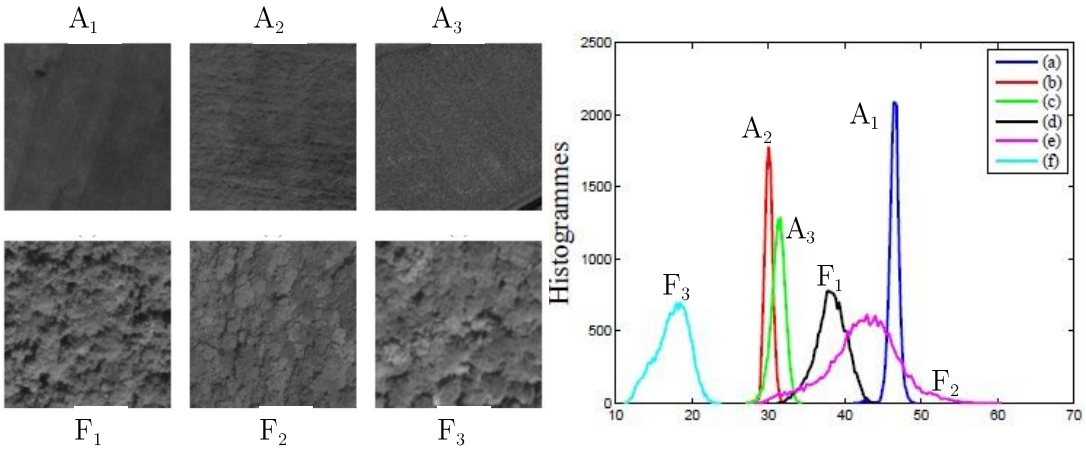


Figure 10 - Textures et histogrammes correspondants

La Figure 10 présente la composante intensité (I) de trois textures de parcelles agricoles (première ligne) et trois textures de forêts (deuxième ligne) avec les histogrammes correspondants. On constate qu'en appliquant directement une analyse sur les intensités, la séparation entre les textures n'est pas évidente puisque les densités des différentes textures se recouvrent.

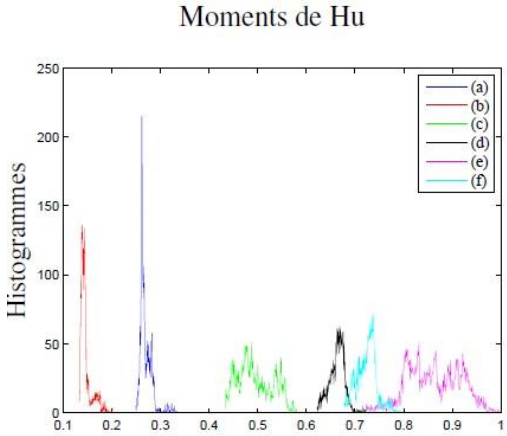


Figure 11 – Histogramme des Moments de Hu calculés sur chaque texture

La Figure 11 montre les histogrammes des moments de Hu calculés sur chaque texture. On constate une meilleure séparabilité des textures en utilisant ce descripteur.

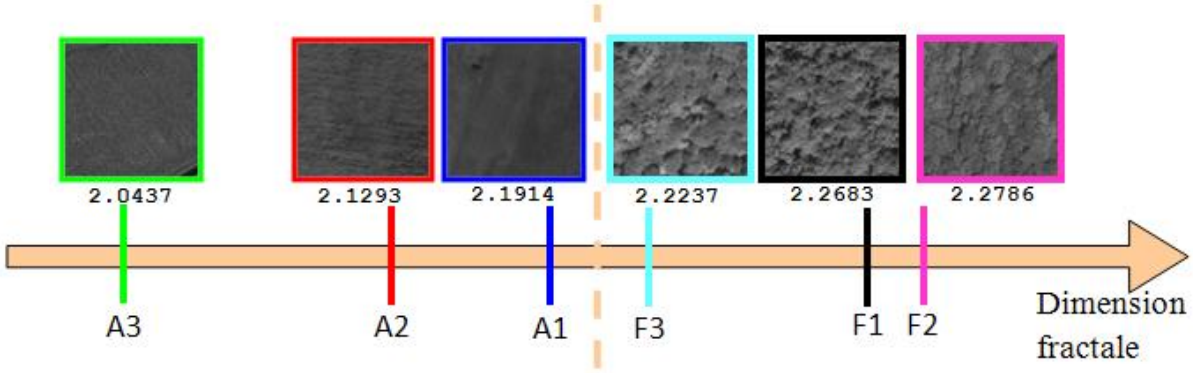


Figure 12 - Dimension fractale

La Figure 12 permet également de définir un seuil sur la dimension fractale pour séparer les textures de forêts des textures agricoles.

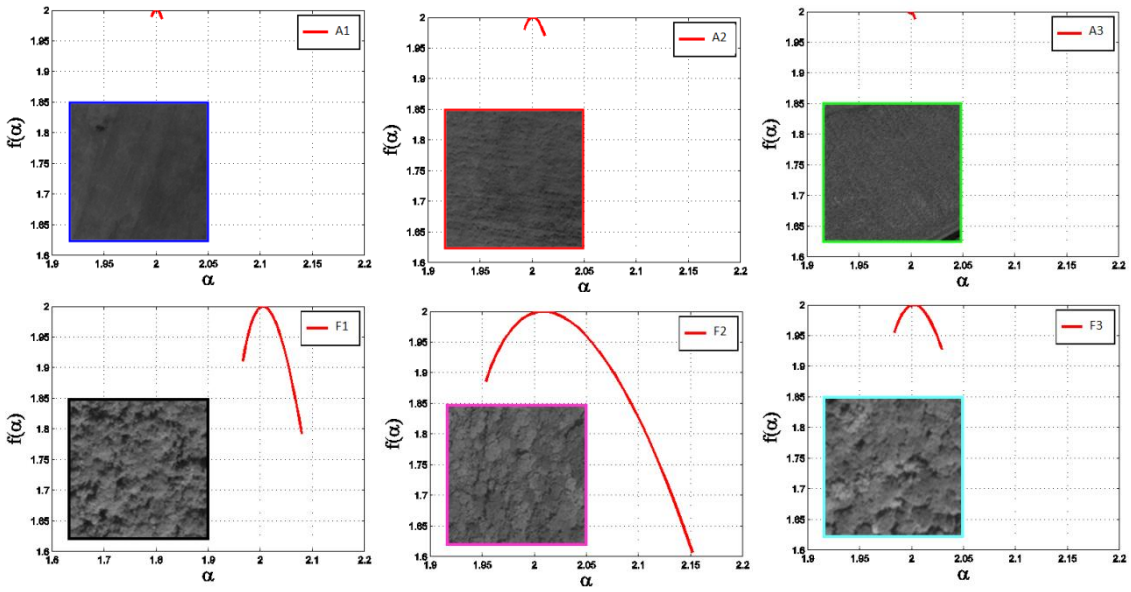


Figure 13 - Spectre de Legendre

Enfin, la Figure 13 illustre le calcul du spectre de Legendre pour les mêmes textures et montre un comportement différent des spectres en fonction de la nature de la texture (forêt ou parcelle agricole).

Ces travaux ont été réalisés essentiellement dans le cadre du stage de DEA de Guianni Commin (2004) et de la thèse de *Mohamed Abadi* (projet CESAR : 2004-2008) et ont donné lieu à 5 publications ([**2**], [**5**], [**17**], [**19**] et [**20**]).

5.2.3 Sélection d'attributs

Même si les descripteurs calculés sur les objets que nous cherchons à caractériser possèdent un pouvoir discriminant intrinsèque, nous montrons dans [**54**] les limites de chacun d'eux pris séparément. Nous montrons également la nette augmentation de la discrimination en combinant plusieurs types de descripteurs (couleur et texture par exemple), ces résultats sont détaillés dans la section 5.2.4 sur la classification. Mais le nombre de descripteurs potentiellement discriminants, que l'on peut calculer sur une image, est très grand (plus de 2000) et la combinatoire trop importante pour que l'on teste toutes les combinaisons possibles. Par ailleurs, les prendre tous en considération simultanément (comme entrées d'un classifieur par exemple) peut poser divers problèmes : *(i)* la mise en œuvre des outils (de classification par exemple) est longue *(ii)* les algorithmes de traitement ne convergent pas toujours *(iii)* les résultats sont souvent peu significatifs du fait de la dispersion des échantillons dans un espace de très grande dimension *(iv)* l'information contenue dans ces descripteurs est redondante, certains d'entre eux étant très corrélés. Afin de réduire la dimension du vecteur décrivant chaque texture (de taille 2184 dans notre cas), nous devons donc sélectionner un sous ensemble pertinent de descripteurs.

Bien qu'initiés pour traiter un problème de classification d'espaces forestiers, ces travaux dépassent largement le cadre applicatif dans lequel ils ont vu le jour pour s'intéresser au concept

plus général de sélection d'*attributs* (ou *variables*) parmi un grand nombre d'attributs. Mes travaux dans ce domaine ont commencé en 2010 avec la signature d'une convention entre le *LAMIA* et le laboratoire *XLIM-SIC* de l'*Université de Poitiers* et ont donné lieu à deux conférences nationales ([**26**], [**28**]), une conférence internationale ([**9**]) une soumission en septembre 2011 dans la revue internationale *Pattern Recognition* ([**39**]).

La sélection d'attributs est utilisée dans de nombreux domaines pour lesquels il existe un grand nombre de variables descriptives d'une entité ou d'un phénomène (biologie, santé, etc.). C'est en effet dans un contexte très général que des critères de sélection et des méthodes de parcours ont été étudiés et proposés. La sélection d'attributs est une étape clé dans plusieurs domaines tels que l'apprentissage, la modélisation de données ou la fouille de données. Elle permet de réduire la dimensionnalité des données en identifiant un ou plusieurs sous-ensembles pertinents parmi l'ensemble des attributs initiaux.

Les travaux menés dans ce cadre ont permis : *(i)* de proposer un critère de sélection basé sur le critère d'information [**9**] pour améliorer les résultats obtenus avec des critères classiques tels que le critère de *Wilk's* [**172**]*(ii)* d'étudier le comportement de la recherche de sous ensembles dans un contexte multi-objectif afin de s'abstraire du problème posé par les critères pour les espaces de très grande dimension ([**26**], [**28**]) *(iv)* de proposer un algorithme hybride basé sur les deux principales approches dans le domaine pour trouver un sous-ensemble optimal au sens de la stabilité dans un contexte de classification [**39**].

Le critère de sélection : critère d'information vs critère de Wilk's

La sélection d'attributs peut se faire selon deux approches principales : l'approche dite *Wrapper* [**129**] qui consiste à appliquer un classifieur sur chacun des sous-ensembles visité pour évaluer sa qualité et les approches dites *Filter* [**174**] qui réalisent une sélection des sous-ensembles à l'aide d'un critère indépendant du classifieur. Nous nous plaçons dans cette dernière catégorie.

Différents critères ont été développés pour évaluer les sous-ensembles candidats. Ils explorent des mesures de distance [**135**] des mesures statistiques [**175**] ou plus récemment des mesures probabilistes basées sur l'estimation de l'information mutuelle [**140**]. Nous avons proposé au cours ne nos travaux d'étudier le potentiel des critères d'information *(IC)* basés sur un calcul du maximum de vraisemblance (*MV*) pénalisé [**183**].

Les critères d'information étudiés sont le critère d'Akaike (AIC [**183**]), le critère d'information bayésien (Bayes Information Criterion : BIC [**149**]) et le critère dit phi-béta (φ_β [**115**]).

$$IC(Y) = -log(MV) + |Y|\alpha(n) \quad (4)$$

où $|Y|$ est le nombre de paramètres libres qui croît en fonction de la complexité du sous-ensemble et $\alpha(n)$ est une fonction pénalisante qui varie en fonction du critère utilisé (respectivement les $\alpha(n)$ de AIC, BIC et φ_β sont égaux à 2, $\log n$ et $n^\beta \log\log n$ avec $0 < \beta < 1$ pour assurer de bonnes propriétés asymptotiques). La quantité $|Y|\alpha(n)$ permet de régulariser le comportement du maximum de vraisemblance car il est connu que ce dernier sur-paramètrise les modèles. Ainsi la minimisation du critère permet d'avoir un compromis entre l'adéquation aux données et la complexité des sous-ensembles candidats.

Pour l'étude de ces critères, les sous-ensembles candidats sont parcourus en utilisant des méthodes standards de la littérature. Plusieurs algorithmes ont en effet été développés pour construire et parcourir les sous-ensembles d'attributs car une recherche exhaustive est très coûteuse en temps de calcul et parfois trop longue pour être envisagée. Les principales méthodes de recherche que nous avons étudié sont les approches séquentielles et en particulier les schémas de sélection progressive (*SFS* pour *Sequential Forward Selection*), rétrograde (*SBS* pour *Sequential Backward Selection*) et flottantes (*SFFS* pour *Sequential Forward Floating Selection* et *SBFS* pour *Sequential Backward Floating Selection*).

Les approches (*IC, Wilk's*) utilisant des critères de sélection différents nous comparons les sous-ensembles retournés en évaluant leurs performances en classification (taux de bonne classification) en appliquant différentes méthodes de classification. Les bases de tests utilisées sont issues des bases de données standards *UCI* telles que *WINE* ou *STATLOG* [**63**].

Nous avons montré dans [**26**] la bonne tenue des critères d'information pour la sélection d'attributs avec des résultats comparables à ceux de la littérature et avec une meilleure stabilité par rapport au critère de *Wilk's* dans la mesure où les différents algorithmes de classification affectent peu notre approche qui converge mieux vers la solution optimale (obtenue par une recherche exhaustive).

Limites de l'approche ascendante et des critères Wilk's et IC

Le principal inconvénient des méthodes ascendantes (*forward*) est leur difficulté à faire émerger des sous-ensembles d'attributs dont l'association permet une bonne description des données. Lorsque plusieurs attributs permettent une bonne description des données en étant réunis mais que pris séparément ils sont peu performants, une approche ascendante aura des difficultés à faire émerger cette association dans la mesure où l'ajout des attributs se fait un à un. Au contraire, une méthode descendante (*backward*) sera plus à même de les conserver dans la mesure où la suppression d'un des attributs concernés réduira considérablement la valeur du critère ce qui devrait écarter ce choix et l'orienter en priorité vers des attributs peu significatifs et/ou sans interactions avec d'autres.

Les approches descendantes sont donc théoriquement plus intéressantes mais leur utilisation avec les critères *IC* et *Wilk's* pose un problème en grande dimension. En effet, lors de l'approche descendante, l'algorithme commence par traiter les espaces les plus grands avant de réduire leur taille. Hors, dans ce cas, non seulement le terme de pénalité des critères *IC* prend très largement le dessus sur le terme de vraisemblance rendant non significatifs les choix de sous ensemble mais pour des tailles de sous-ensemble trop grandes, le calcul même du critère est remis en cause car le calcul du critère d'information pour de grandes dimensions pose le problème de l'estimation de densités de probabilité multi-variées en grande dimension.

Cependant nous présentons des pistes possibles dans les perspectives (section 6 page 82) pour appréhender ce problème.

Approche multi-objectif et information mutuelle

Cette approche a été mise en place pour résoudre le problème soulevé précédemment. Ce problème étant toujours ouvert nous avons décidé d'utiliser une approche multi-objectif ascendante en optimisant la *pertinence* D et la *redondance* R de l'information contenue dans les attributs sélectionnés et calculées à partir de l'*information mutuelle* (*IM*).

Soient X et Y deux variables aléatoires, l'information mutuelle $IM(X;Y)$ est définie par $P(X)$, $P(Y)$ et $P(X,Y)$ (lois de probabilité discrètes)

$$(5) \qquad IM(X;Y) = \sum_{y\in Y}\sum_{x\in X} P(x,y)\cdot \log\frac{P(x,y)}{P(x)\cdot P(y)}$$

$IM(X;Y)$ est élevée si X et Y sont dépendantes. Dans le cas où $IM(X;Y)=0$ cela signifie que les variables sont indépendantes. La pertinence D est choisie comme la valeur maximale de IM entre les variables prises individuellement et les classes (c).

$$(6) \qquad D = \max_{1,\cdots,M}\frac{1}{|S|}\sum_{X_i\in S} IM(X_i;c)$$

Il a été reconnu que la sélection des meilleures variables, en utilisant uniquement D ne retourne pas nécessairement la solution optimale ([**56**], [**120**]) car les variables peuvent être redondantes entre elles. Différents moyens existent pour vérifier la redondance entre les variables ([**91**], [**120**]). Celle-ci peut être définie aussi par l'information mutuelle entre deux variables $(X_i;X_j)_{\substack{i,j=1,\cdots,m\\ i\neq j}}$ appartenant à l'ensemble sélectionné S de m variables.

$$(7) \qquad R = \min_{1,\cdots,M}\frac{1}{|s|^2}\sum_{X_i,X_j\in S} IM(X_i;X_j)$$

Comme dans l'étude précédente, l'approche proposée est de type *filter* et nous comparons les résultats des différentes méthodes de sélection en appliquant des classifieurs (*Naive Bayes NB, K-Nearest Neighbor KNN*, *Linear Discriminant Analysis LDA*). Nous présentons dans le Tableau 3 une comparaison quantitative sur des bases de données *UCI* [**63**] entre les résultats obtenus par les algorithmes *mRMR (minRedundancy-MaxRelevance)* [**171**], *FSDD* (Feature Selection algorithm based on a Distance Discriminant) [**135**] et par notre algorithme *mRMR-PC (min Redundancy Max Relevance Pareto Curve)*. Le Tableau 3 montre les moyennes (μ) de taux de bonne classification, leur écarts-type (σ) et le nombre de variables (#) qui constituent la combinaison possédant le meilleur taux de bonne classification. Nous remarquons la stabilité et les performances de notre approche sur tous les jeux de données et pour les quatre classifieurs. Nous constatons aussi que si le nombre de variables est petit, notre approche converge vers la solution optimale obtenue par une recherche exhaustive[1] (ex. jeux de données IRIS) contrairement aux autres. Il ressort de cette étude que notre approche converge plus rapidement [**9**] vers l'optimum global.

Approche multi-objectif hybride

Afin d'améliorer encore les résultats nous intégrons la phase de validation précédente (évaluation des classifieurs) pour réaliser la sélection des sous ensembles les plus stables parmi ceux retournés par la phase *filter*.

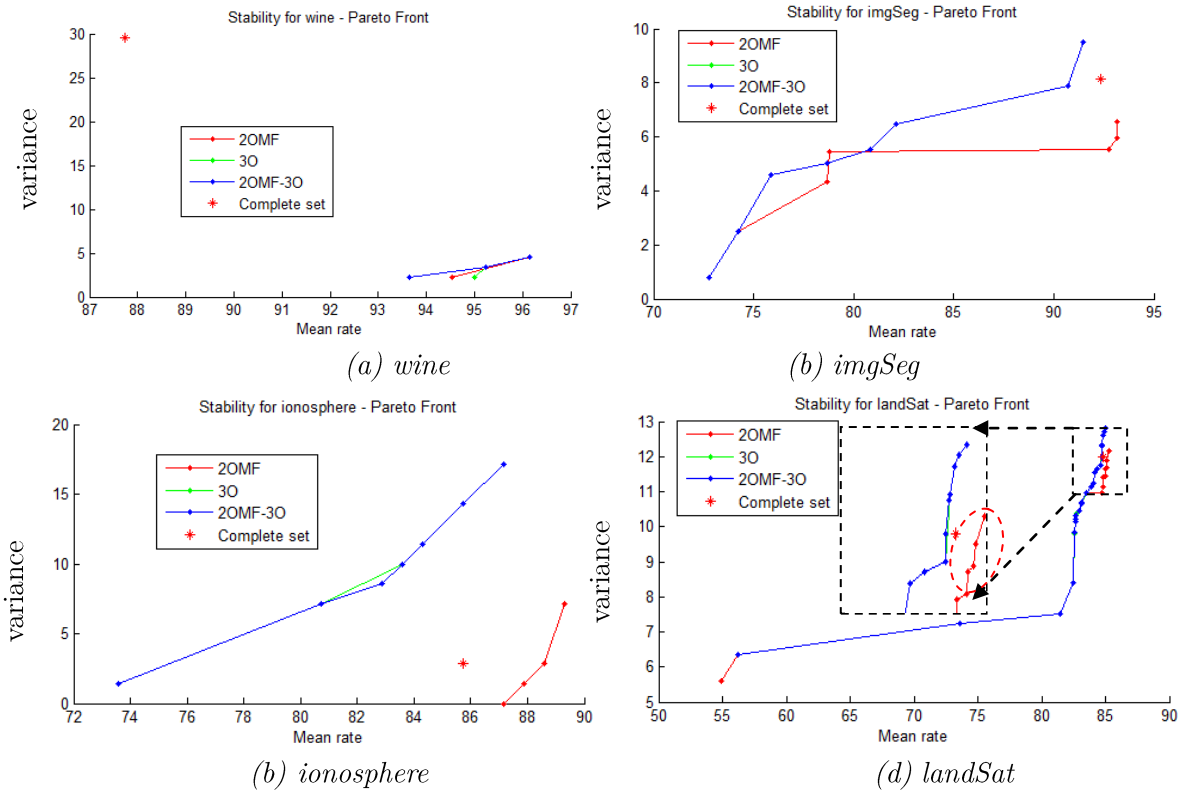
[1] Toutes les combinaisons possibles. Il existe 2^N-1 combinaisons (N est le nombre de variables).

Jeux de Données			IRIS	TAE	WINE	SATLOG1	SATLOG2	SPAMBASE	MFEAT (50)	Moyenne
mRMR+PC	**kNN**	μ	97.43	**64.28**	80.32	95.12	**90.05**	**90.82**	89.90	**86.846**
		σ	0.56	1.81	1.26	0.46	--	--	--	
		#	2	2	12	16	25	54	47	
	LDA	μ	**98.53**	54.77	**99.05**	91.30	84.05	89.83	**98.50**	88.004
		σ	0.66	1.66	0.44	0.24	--	--	--	
		#	2	3	12	18	35	48	47	
	NB	μ	96.62	55.15	98.96	87.82	78.70	89.91	97.20	86.337
		σ	0.60	1.02	0.45	0.41	--	--	--	
		#	2	4	12	17	32	52	29	
mRMR	**kNN**	μ	95.92	60.37	78.33	95.12	**90.05**	90.43	84.70	84.998
		σ	0.22	2.04	1.37	0.46	--	--	--	
		#	4	5	8	16	28	53	48	
	LDA	μ	97.97	53.79	99.02	91.30	84.00	89.87	**98.50**	87.778
		σ	0.16	0.85	0.46	0.24	--	--	--	
		#	4	4	12	18	34	53	47	
	NB	μ	96.00	54.97	95.65	87.39	79.20	89.30	97.20	85.673
		σ	0.00	0.91	0.54	0.57	--	--	--	
		#	4	4	6	18	27	56	29	
FSDD	**kNN**	μ	96.61	60.37	77.70	**95.21**	89.85	90.48	89.80	85.717
		σ	0.69	2.04	1.19	0.47	--	---	--	
		#	2	5	11	12	27	43	47	
	LDA	μ	97.97	53.98	98.81	91.30	84.05	89.48	98.30	87.698
		σ	0.16	0.78	0.47	0.24	--	--	--	
		#	4	4	13	18	35	57	48	
	NB	μ	96.00	54.83	95.45	87.43	79.20	89.35	95.70	85.423
		σ	0.23	1.07	0.34	0.58	--	--	--	
		#	3	5	13	17	27	53	32	

Tableau 3 - Comparaison de l'approche multi-objectif avec d'autres approches

Pour cela nous avons défini un algorithme hybride [**39**] basé sur une première étape de type *filter* avec comme critères la pertinence et la redondance de l'information exprimés précédemment. Cette première étape permet de fournir un pseudo-ensemble de Pareto en construisant des fronts de Pareto successifs pour chacune des tailles d'espace possibles. En effet, après avoir testé plusieurs approches, c'est cet algorithme, appelé *2OMF (2 Objectifs Multi Front)*, qui donne les meilleurs résultats en terme de performance et de diversité des solutions [**39**]. Dans une deuxième étape, nous appliquons donc un ensemble de classifieurs pour évaluer chacun des sous-ensembles conservés. Finalement l'algorithme renvoie les sous-ensembles non dominés vis-à-vis des deux critères suivants : maximisation du taux moyen de bonne classification obtenu sur les classifieurs et minimisation de la variance du taux.

La Figure 14 montre les résultats obtenus sur la base *UCI* [**63**] en affichant les fronts de Pareto obtenus après la deuxième étape. On constate une large domination des fronts issus de la méthode *2OMF* (en rouge) par rapport aux autres méthodes et par rapport à l'ensemble complet d'attributs (étoile rouge).



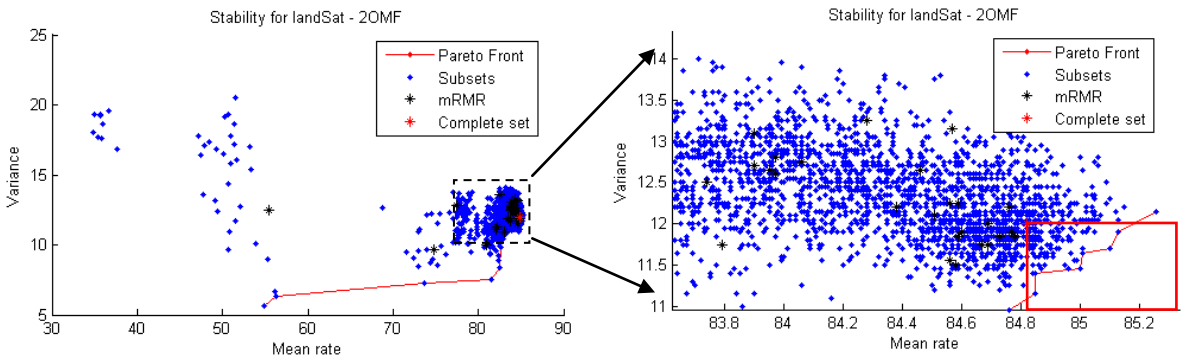
(a) wine *(b) imgSeg*

(b) ionosphere *(d) landSat*

Figure 14 - Comparaison des stabilités pour différentes bases (a) wine (b) imgSeg (c) ionosphere (d) landSat

La raison principale du succès de *2OMF* est que l'exploration systématique de toutes les tailles de sous-ensemble combinée à l'approche multi-objectif permet de conserver une grande diversité dans les solutions conservées et ainsi de contourner la limitation classique des approches ascendantes qui échouent en tombant dans un optimum local dont elles ne peuvent sortir qu'en ajoutant plusieurs attributs à l'ensemble courant et en particulier pour les attributs donnant une bonne description des données lorsqu'ils sont associés à d'autres attributs.

La Figure 15 affiche la stabilité dans le cas des données LandSat. La Figure 15 (b) (qui est un zoom de la Figure 15 (a)), montre que beaucoup de sous-ensembles (ceux à l'intérieur du rectangle rouge), parmi ceux sélectionnés après l'étape *filter*, dominent la stabilité de l'ensemble complet (étoile rouge). Certains de ces sous-ensembles sont Pareto optimaux et d'autres ont une taille très faible comparée à l'ensemble complet. Par ailleurs, dans cette même figure sont affichés (étoiles noires) les différents sous ensembles parcourus avec l'algorithme *mRMR* [**171**]. On constate que nous seulement les sous-ensembles ne sont pas Pareto optimaux mais qu'en plus dans le cas *landSat* aucun d'eux ne domine le sous ensemble complet d'un point de vue stabilité. Ces tendances se retrouvent dans les autres bases de données testées avec parfois quelques sous-ensembles issus de *mRMR* dominant le sous ensemble complet mais, à l'exception de la base *TAE*, aucun d'eux n'est présent sur la courbe de Pareto. Notre approche permet donc d'obtenir des solutions dominant à la fois le sous ensemble complet et les solutions renvoyées par l'algorithme *mRMR* pour une complexité de calcul raisonnable même si l'algorithme *mRMR* est plus rapide.



(a) Front de Pareto (b) Sous ensembles dominants

Figure 15 - Analyse de la stabilité pour la base landSat : (a) Front de Pareto (b) Sous ensembles dominants

Dans [**39**], nous montrons que beaucoup des sous ensembles dominant la solution complète possèdent entre 30% et 60% d'attributs en moins ce qui est fort intéressant et qui prouve qu'une réduction conséquente du nombre d'attributs peut améliorer les performances de la classification (nous ajoutons aux classifieurs précédemment utilisés les classifieurs *Probabilistic Neural Network PNN* et *Mahalanobis Mah*). Nous remarquons également que les plus petites tailles de sous ensembles qui dominent l'ensemble complet ne font pas partie de l'ensemble de Pareto.

Le Tableau 4 montre des sous-ensembles issus du front de Pareto ($PO = Y$) et aussi des sous-ensembles en dehors du front de Pareto ($PO = N$) mais qui dominent le sous-ensemble complet. Toutes ces solutions sont intéressantes car elles possèdent un faible nombre d'attributs et une meilleure stabilité.

				Taux						
Bases	Sous ensemble	Taille	PO	Mean	Var.	KNN	LDA	Mah	NB	PNN
Wine	Ensemble complet	13	N	87.7	29.6	70.4	**100**	92.0	96.6	79.5
	[1 7 11]	**3**	**Y**	**96.2**	**4.5**	**97.7**	93.2	**96.6**	**97.7**	**95.4**
	[1, 2, 6, 7, 8, 11, 12]	**7**	**Y**	**94.5**	**2.2**	**95.5**	93.2	**93.2**	**95.5**	**95.5**
	[1 7]	**2**	N	**94.3**	**5.6**	**96.6**	90.9	**94.3**	**95.5**	**94.3**
imgSeg	Ensemble complet	18	N	92.3	8.2	95.5	**91.1**	NA	87.3	95.3
	[2 10 12 13 15 17 18]	**7**	**Y**	**93.2**	**6.5**	**96.1**	90.9	NA	**89.5**	**96.0**
	[2 9 10 12 13 15 17 18]	**8**	**Y**	**93.1**	**5.9**	**96.0**	90.7	NA	**90.0**	**95.6**
	[2 5 10 11 12 13 15 17 18]	**9**	**Y**	**92.7**	**5.5**	95.2	90.3	NA	**89.8**	**95.4**
ionoSphere	Ensemble complet	34	N	85.7	2.8	87.1	NA	NA	NA	84.2
	[...]	**32**	**Y**	**87.1**	**0**	**87.1**	NA	NA	NA	**87.1**
	[...]	**21**	N	**87.8**	**1.4**	**88.6**	NA	NA	NA	**87.1**
	[...]	**22**	N	**87.8**	**1.4**	**88.6**	NA	NA	NA	**87.1**
landsat	Ensemble complet	36	N	84.8	12	89.4	83.9	81.6	78.5	90.5
	[...]	**29**	**Y**	**85.0**	**11.4**	**89.8**	83.7	**82.1**	**78.9**	90.3
	[...]	**31**	**Y**	**85.1**	11.9	89.6	83.7	**83.0**	**78.7**	**90.6**
	[...]	**25**	N	**84.9**	**11.9**	89.3	83.2	**82.7**	**78.7**	**90.6**
	[...]	**26**	N	**84.8**	**11.7**	**89.6**	83.2	**82.3**	**78.6**	**90.6**

Tableau 4 - Taux de bonne classification et stabilité pour quelques sous-ensembles intéressants

Par exemple, pour la base *wine* un sous-ensemble à 2 attributs (2 et 7) est plus performant que l'ensemble à 13 attributs (meilleurs taux de bonne classification pour 4 classifieurs sur 5). De la même manière pour la base *imgSeg* le nombre d'attributs est divisé par 2. Nous avons donc avec cette méthode *2OMF* une approche permettant de réduire considérablement la taille des ensembles d'attributs.

5.2.4 Classification d'images satellites

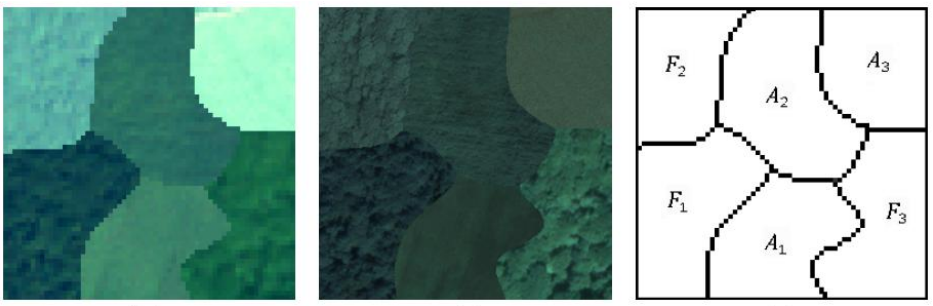
Cette étape est la dernière de la première partie de la chaîne. La thématique a démarré dès 2004 avec le stage de DEA de *Guianni Commin* sur la *caractérisation de textures forestières* [**79**]. Ce stage a permis de trier un certain nombre de techniques de caractérisation en fonction de leur pertinence (Moments statistiques, Matrices de Cooccurrence, champs de Markov, …) qui seront par la suite repris dans le cadre du projet *CESAR* et de la thèse de *Mohamed Abadi* (2004-2008) [**54**] et lors des stages de doctorat de *Saliha Loumi et Farid Alilat* (2007) [**37**].

La principale contribution sur cette thématique a été apportée lors de la thèse de *Mohamed Abadi* (soutenue en 2008). L'analyse de l'information spectrale et structurelle des images satellites à très haute résolution spatiale (inférieur à un mètre), permet d'obtenir une grande précision en termes de reconnaissance et localisation (à l'arbre près dans certains contextes). Les travaux menés ont permis d'obtenir des outils de classification basés sur l'information spectrale, structurelle et fractale afin de cartographier les espaces forestiers étudiés. Ces outils utilisent la sélection d'attributs détaillée précédemment afin de permettent une meilleure séparation des classes ainsi que les techniques de changement d'espace couleur et de fusion d'images présentées également précédemment et permettant d'utiliser conjointement une information spectrale riche et une information spatiale précise.

Nous présentons ici les principaux résultats obtenus dans ce cadre, les publications relatives à ces travaux étant ([**2**], [**5**], [**13**], [**16**], [**17**], [**18**], [**19**] et [**54**]).

Les principaux algorithmes de classification retenus sont les *Modèles de Mélange de Gaussiennes* (*MMG*) ([**79**], [**199**]), les *Support Vector Machine* (*SVM*) [**202**], les *K-Means* et la méthode *ISODATA* [**62**]. Ils ont été sélectionnés pour disposer d'un ensemble de classifieurs ayant des caractéristiques différentes afin de mesurer l'impact du classifieur sur notre approche.

L'apprentissage et la validation des méthodes ont été effectués sur des données de simulation (exemple donné dans la Figure 16) ainsi que sur des données réelles (exemple donné dans la Figure 17).



(a) (b) (c)

Figure 16 - Exemple de données de synthèse pour la validation des classifieurs (a) image basse résolution, (b) image haute résolution (c) répartition des classes



I_1 I_2 I_3 I_4

Figure 17 - Exemple de données réelles (classées par des experts) pour la validation des classifieurs

La Figure 18 et la Figure 19 présentent les résultats obtenus pour les différents classifieurs. Même si des différences existent entre les résultats, les taux de bonne classification (Tableau 5) pour les images de synthèse sont relativement bons compte tenu de la complexité des classes (objets texturés). L'expression du calcul de la sensibilité (Sp) et de l'Erreur Globale (Eg) sont extrait de [**137**].

Classifieur	Taux de bonne classification (Sensibilité Sp)	Erreur Globale (Eg)
K-Means	94.78	9.79
ISODATA	96.39	22.64
MMG	97.21	8.11
SVM	97.12	18.52

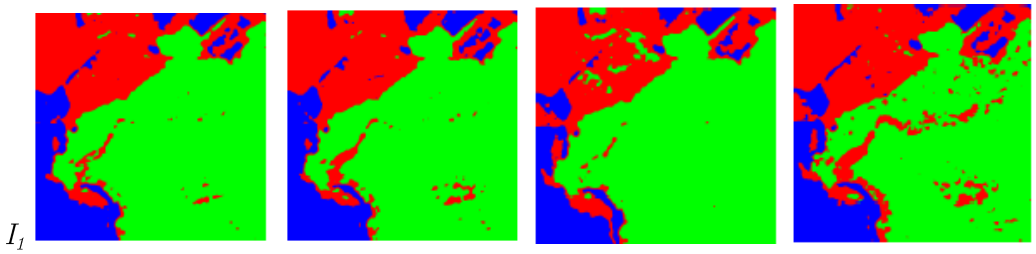
Tableau 5 - Taux de bonne classification et Erreur Globale pour les données de synthèse



KMeans *Isodata* *MMG* *SVM*

Figure 18 – Résultats de la classification des images de synthèse

Concernant les images réelles, l'examen visuel par des biologistes (indiquée dans la Figure 17) a permis de conclure à de très bon taux de bonne classification.



I_1

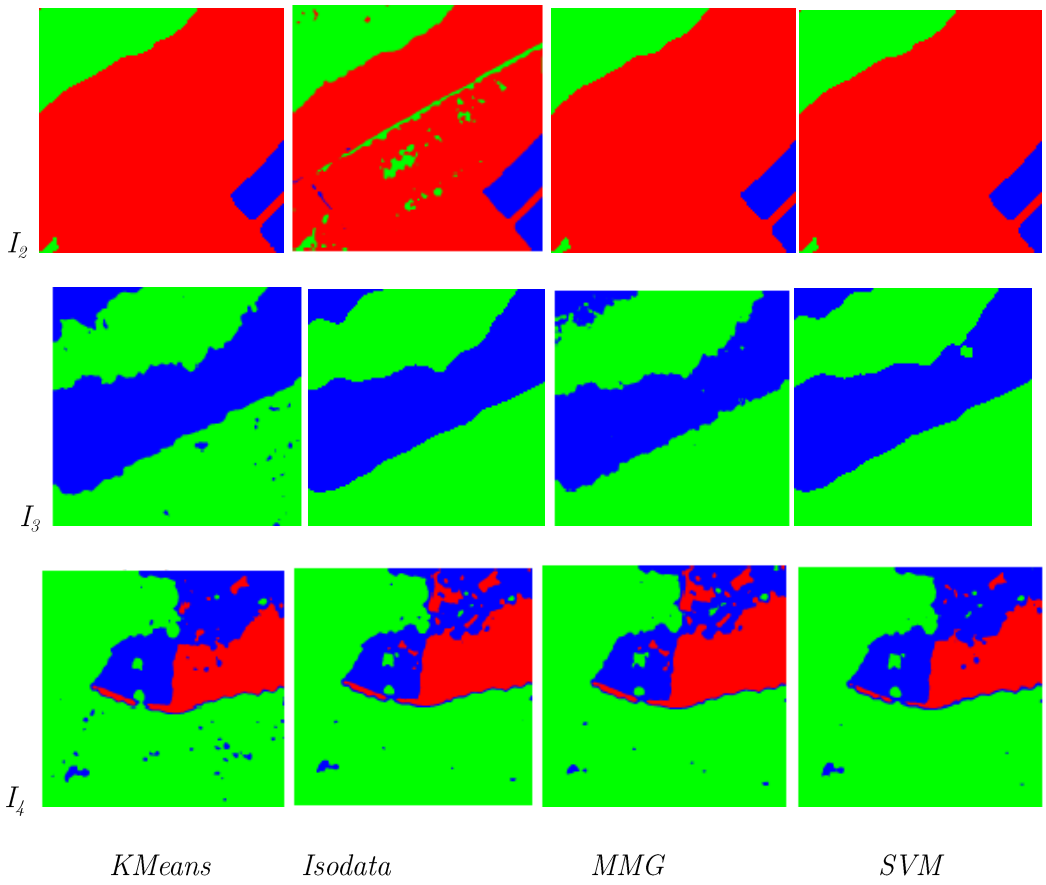


Figure 19 - Résultats de la classification des images naturelles

Ces travaux ont par ailleurs été valorisés lors du stage de Master 2ème année d'*Ikram El-Missi* (2007) [**95**] dans le cadre du projet *PARAGE* (occuPation Agricole dans les Régions Antilles et Guyane) regroupant Spot Image, le *CIRAD*, l'*IRD* et *SIGBEA* [**173**]. Les outils de classification développés, et notamment la caractérisation des textures, ont permis de montrer l'apport de la télédétection pour la surveillance des parcelles agricoles. Les outils ont été utilisés afin de séparer les parcelles agricoles des zones forestières et pour différencier des parcelles agricoles entre elles. Des développements spécifiques ont été effectués avec notamment l'utilisation de seuillages d'histogrammes ou d'ondelettes analysantes (*Lorentzienne* et *Gaussienne*). La Figure 20 présente des résultats de séparation forêt agriculture dans le cas d'images *Spot5*. La Figure 21 présente des résultats de séparation dans le cas d'une scène étendue *QuickBird*.

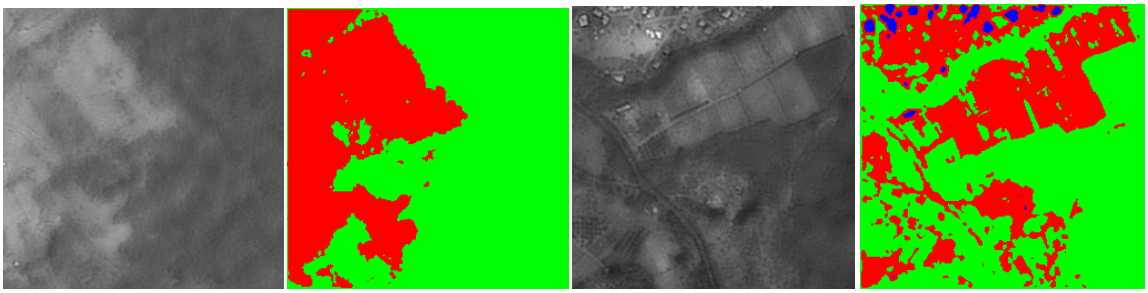


Figure 20 - Séparation forêt agriculture sur des images Spot 5

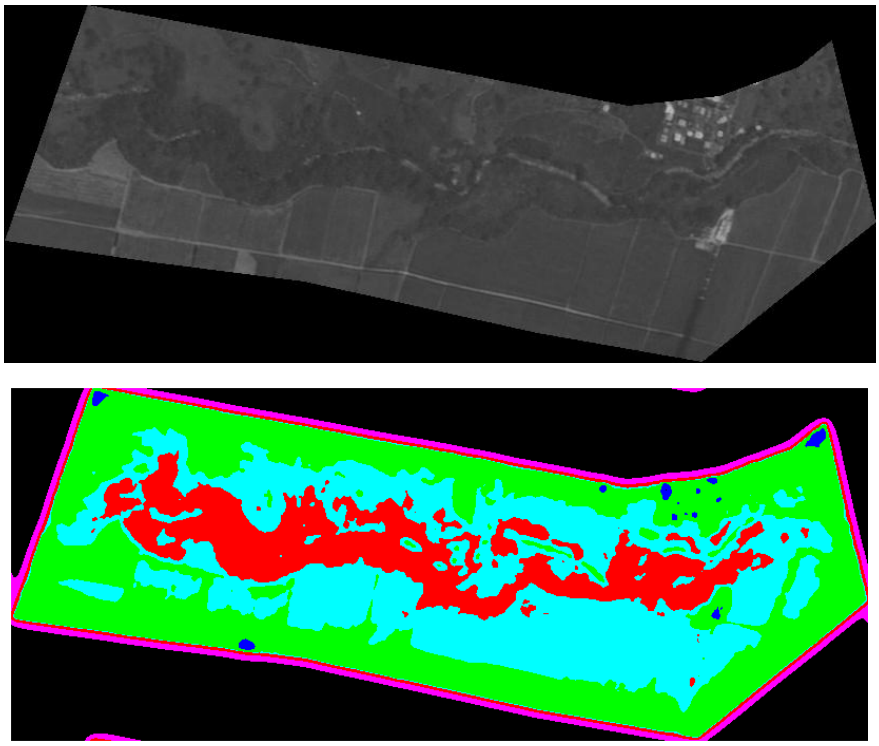


Figure 21 - Séparation forêt agriculture sur une scène Quickbird

Les limites de ces travaux résident dans l'application des algorithmes de classification sur de larges scènes comportant un trop grand nombre de classes, la séparation devenant beaucoup plus délicate. Afin de repousser les limites de ces méthodes, il est nécessaire d'introduire des connaissances supplémentaires sur les scènes traitées. C'est l'objectif de la seconde partie de la chaîne de traitement qui présente des modèles et des approches complémentaires permettant de réaliser la classification de couverts forestiers sur de plus vastes zones (l'ensemble de la *Basse-Terre* dans notre cas).

5.3 Modélisation des données

Comme indiqué précédemment, les classifications ne faisant intervenir que des informations provenant des images satellites trouvent des limites lors de leur application à grande échelle. Nous avons donc abordé le problème du point de vue de la *sémantique de l'information* afin de modéliser les scènes de notre région d'intérêt et ainsi séparer le problème global de classification en sous problèmes pouvant être résolus d'un point de vue purement sémantique ou en utilisant les techniques précédentes (stratégie diviser-pour-régner).

La première étape concerne la constitution d'un dictionnaire sur les couverts forestiers intégrant des données sémantiques et images (5.3.1) afin de mieux connaître les entités que l'on cherche à localiser. Les rapprochements établis avec le laboratoire *DYNECAR* et notamment avec *Alain Rousteau* ont permis d'aborder le problème de la classification des espaces forestiers par une approche vectorielle faisant intervenir des critères environnementaux et topologiques, qui sont en partie responsables des caractéristiques des formations forestières, plutôt que de se focaliser sur leur aspect à travers les images satellites (5.3.2). Cette technique est ensuite étendue à un autre

problème concernant la classification des habitats dans le cadre de la propagation du virus de la dengue (5.3.2.2).

Ces travaux ont permis de mieux connaître les phénomènes que nous cherchions à classer et à mieux comprendre les limites des outils que nous utilisions. En effet, le caractère transitoire des formations forestières induit un gradient dans les changements de valeurs des attributs calculés (textures) rendant difficile la séparation des classes. Ces études ont soulevé la problématique plus générale de la classification et de la représentation vectorielle de données aux frontières diffuses pour laquelle nous proposons une approche et un modèle (5.3.3). En effet, les structures de donnée vectorielles présentes dans les *SIG* sont adaptées à la modélisation d'ensembles stricts mais pas à la modélisation d'ensembles flous. Rappelons que l'intérêt pour le format vectoriel fait suite à la constatation que l'utilisation des données brutes *raster* (image) au sein des *SIG* n'est pas envisageable, dans [**65**] les auteurs soulignent l'importance du passage des images vers les *SIG* de manière à disposer d'une *information exploitable* dans un format standard. Ils précisent que les données raster ne sont pas adaptées à une manipulation dans un environnement de *SIG* (manque de souplesse, volume important des données) et qu'elles ne permettent pas d'intégrer une information contextuelle.

Enfin, les différentes approches menées ont fait apparaître le problème de la mise en correspondance de différentes représentations d'une même entité lors du croisement de couches d'information. En effet, outre la précision du positionnement des sommets, la vectorisation d'un objet réel engendre un certain nombre de problèmes intrinsèques au procédé lors du choix du nombre de sommets (lié à la résolution) et à leur emplacement (très souvent arbitraire). Ceci conduit à des approximations différentes d'une même entité sans pour autant qu'une soit plus juste qu'une autre ([**70**], [**110**]). Il faut rappeler que les données géographiques contenues dans un *SIG* représentent un modèle conceptuel de la réalité et qu'il peut en exister plusieurs sans que des outils soient proposés pour les mettre en correspondance. Pour résoudre le problème, nous proposons un modèle sémantique et relationnel des couches d'information (5.3.4).

5.3.1 Le dictionnaire des espaces forestiers de la Guadeloupe

Les différentes caractérisations des forêts au travers de l'analyse de leur couleurs et textures par de multiples descripteurs ainsi que les groupements écologiques définis par *Alain Rousteau* en 1996 [**181**] nous permettent de définir un dictionnaire des espaces forestiers formalisés au sein d'une ontologie (Figure 22) (projet de master 2ème année 2010-2011 de *Gary Poinin* et *Hulrick Kodaday*).

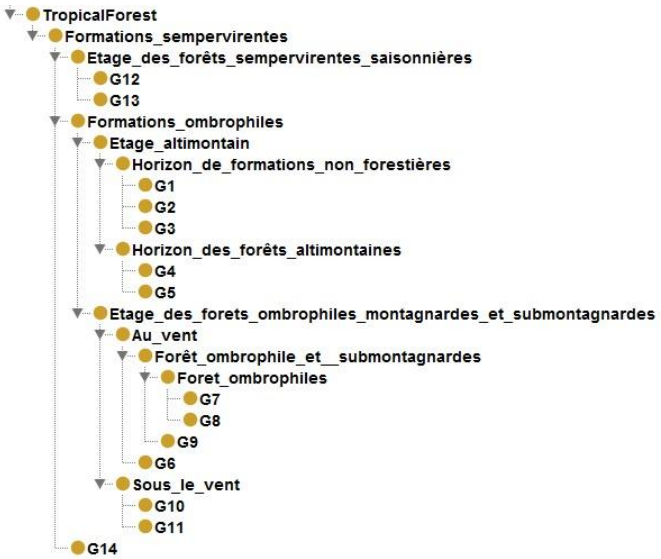
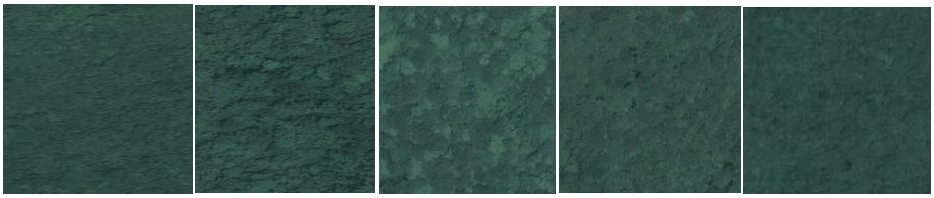


Figure 22 - Ontologie des couverts forestiers de la Guadeloupe

Chacun des groupements écologiques matérialisé par un concept dans l'ontologie est renseigné par : *(i)* des informations sur le climat (température, précipitations, etc.), certains caractères des sols, la structure et la composition des formations forestières, la biodiversité ou les catégories de risques encourus par les écosystèmes *(ii)* une base de textures (Figure 23) extraites de différentes images (prises par des capteurs différents et à des dates différentes). Chacune de ces textures est renseignée par un ensemble de descripteurs sélectionnés parmi les 2184 descripteurs précédents *(iii)* une estimation des informations primaires sur la canopée (surface moyenne des couronnes, etc.) et les arbres (densité, nombre, etc.) obtenue par notre partenaire Canadien de l'Université de Moncton, *Eric Hervet* (Figure 24) [**98**].

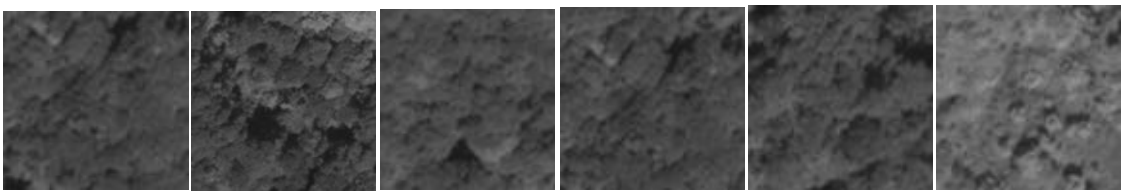
Etage Altimontain (Images Multi-spectrales 2.4 m Quickbird)



Forêts ombrophiles au vent (Images Multi-spectrales 4m et Panchromatiques 1m IKONOS)



Forêts ombrophiles sous le vent (Images Panchromatiques IKONOS, 1m)



Formations sempervirentes (Images Multi-spectrales Quickbird 2.4m, Spot5 5m, IKONOS 1m)

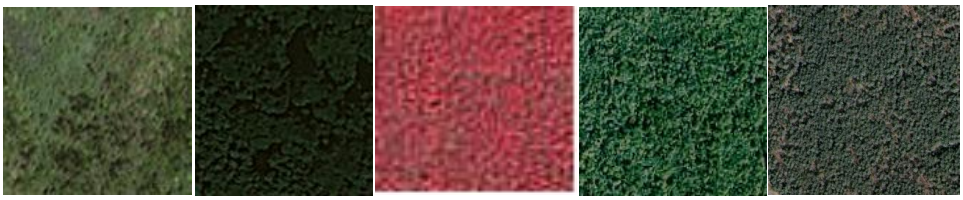


Figure 23 - Extraits de la base de textures

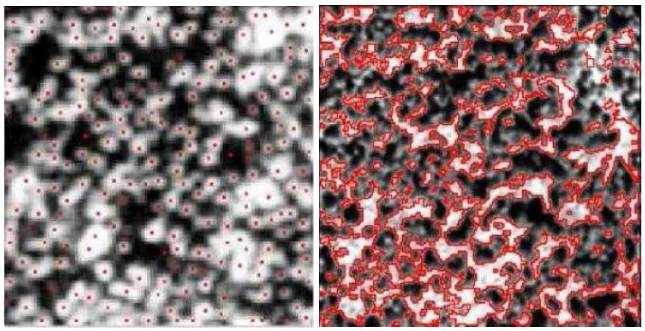


Figure 24 - Localisation des cimes et des couronnes des arbres

La base d'images de textures est renseignée à l'aide de métadonnées descriptives [**187**] reprenant l'ensemble des éléments cités précédemment. Ces métadonnées sont codées en *OWL* (***O**ntology **W**eb **L**anguage*), langage basé sur *XML* (*e**X**tensible **M**arkup **L**anguage*) en suivant le modèle *RDF* (***R**esource **D**escription **F**ramework*).

Ce dictionnaire est en cours de finalisation.

5.3.2 Classification par approche vecteur

Nous nous intéressons maintenant à la classification à partir de données vectorielles, c'est-à-dire provenant de couches d'information issues d'un *SIG,* et non plus de données raster, issues des images. La principale différence est que la classification va être basée sur la sémantique de l'information (attributs numériques et/ou symboliques) et qu'elle sera appliquée sur des objets vectoriels issus du croisement d'un certain nombre de couches d'information et non plus sur des pixels d'une image. Les méthodes de classifications utilisées seront principalement des arbres de décision [**75**].

5.3.2.1 Mise à jour de la cartographie des espaces forestiers de la Guadeloupe

Les travaux menés sur la classification d'images satellites (5.2.4) ont pour principale application la classification des espaces forestiers de la Guadeloupe avec comme objectif de mettre

à jour la cartographie (Figure 25) obtenue en 1996 par *Alain Rousteau* [**181**]. Mais comme souligné à plusieurs reprises, cette mise à jour par télédétection uniquement trouve des limites en raison de la densité et de la complexité des formations forestières. Pour aborder le problème de manière sémantique, nous avons eu recours à une expertise (laboratoire *DYNECAR*) car la définition même des classes représentant les formations forestières est sujette à discussion pour diverses raisons : manque de relevés terrain dans certaines zones complexité des formations en termes de composition floristique et ambiguïté sur les critères floristiques de qualification des formations. Ces connaissances expertes ont été en partie formalisées grâce au dictionnaire précédent.

L'intérêt de l'approche vecteur est donc de pouvoir traiter l'ensemble de la zone d'intérêt et d'intégrer directement les résultats au sein d'un *SIG*. La carte écologique de la Guadeloupe représente un état des connaissances en matière d'inventaire écologique et de traitement de l'information. Ce document a permis d'identifier pour la seule Guadeloupe, 38 unités écologiques naturelles (Figure 25) et 6 modalités d'artificialisation dans les régions transformées par l'homme. Cette carte écologique rend des services pratiques en matière de gestion des espaces ; elle contribue à diffuser la connaissance sur les systèmes écologiques caribéens et offre finalement un cadre pour étudier, gérer et conserver le patrimoine naturel antillais. Elle mérite cependant d'être complétée et affinée.

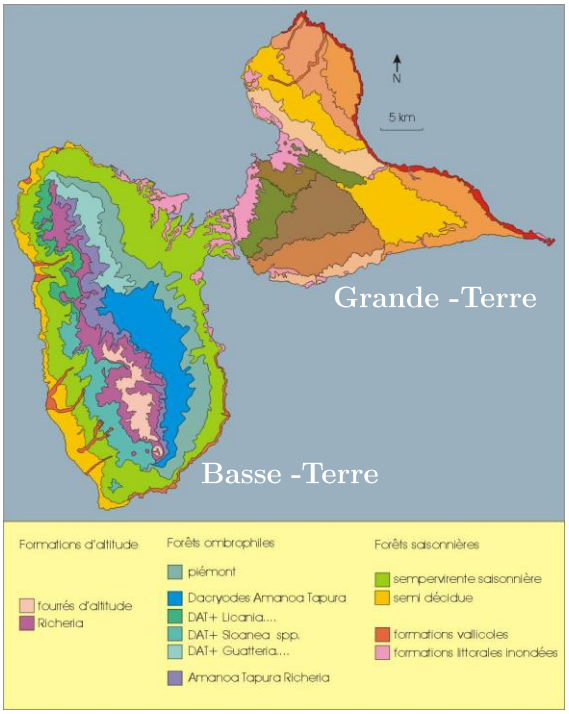


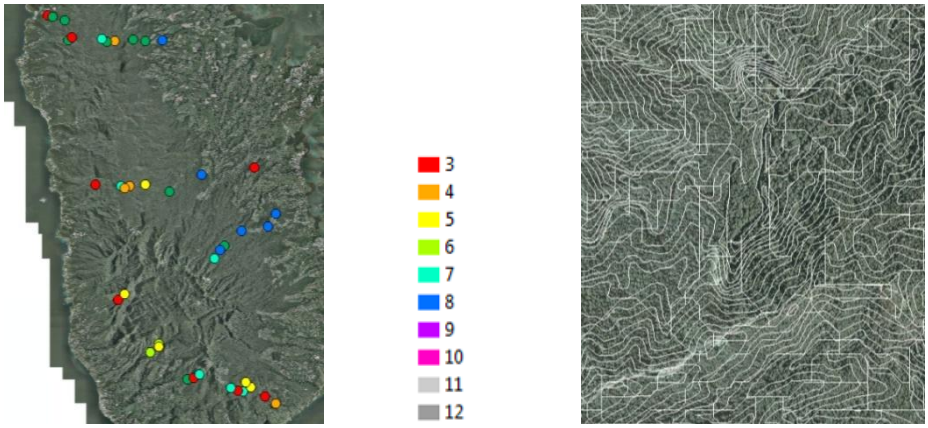
Figure 25 - Cartographie des unités écologiques naturelles de la Guadeloupe. (La légende ne référence que les unités représentées en Basse-Terre [181])

Cette cartographie a été obtenue de la manière suivante : *(i)* les classes ont été définies à partir de l'information floristique contenue dans les relevés effectués manuellement sur une cinquantaine de parcelles réparties sur la Basse-Terre (Figure 26 (a)) *(ii)* la localisation des classes a été construite manuellement en tenant compte de la localisation des parcelles et d'informations topographiques telles que l'altitude, la pente ou l'exposition.

Nous avons dans un premier temps automatisé la procédure de classification de manière à augmenter la précision de la cartographie sans remettre en question la taxonomie des classes proposée par les biologistes en se basant uniquement sur l'information floristique.

Pour cela, nous avons :

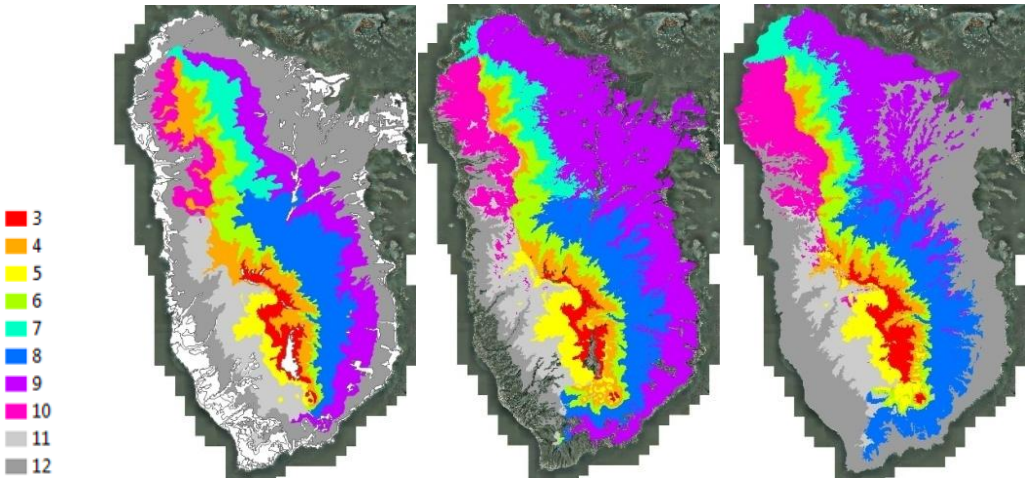
1. fusionné les différentes couches d'information afin d'obtenir une partition de la Basse-Terre en zones (polygones) homogènes du point de vue des couches thématiques fusionnées (pente, altitude, versant, exposition) (de l'ordre de 410 000 unités obtenues, Figure 26 (b)). On obtient ainsi un ensemble d'unités appelées *UEV* pour ***U**nités **E**cologiques **V**ectorielles*.
2. réalisé un apprentissage sur les données topographiques (pente, altitude, versant, exposition et latitude) extraites des parcelles classées.
3. appliqué une classification supervisée sur l'ensemble des unités afin de les labelliser.



(a) (b)

Figure 26 – (a) Données d'apprentissage (Placettes d'observation) et (b) Extrait de la carte des unités

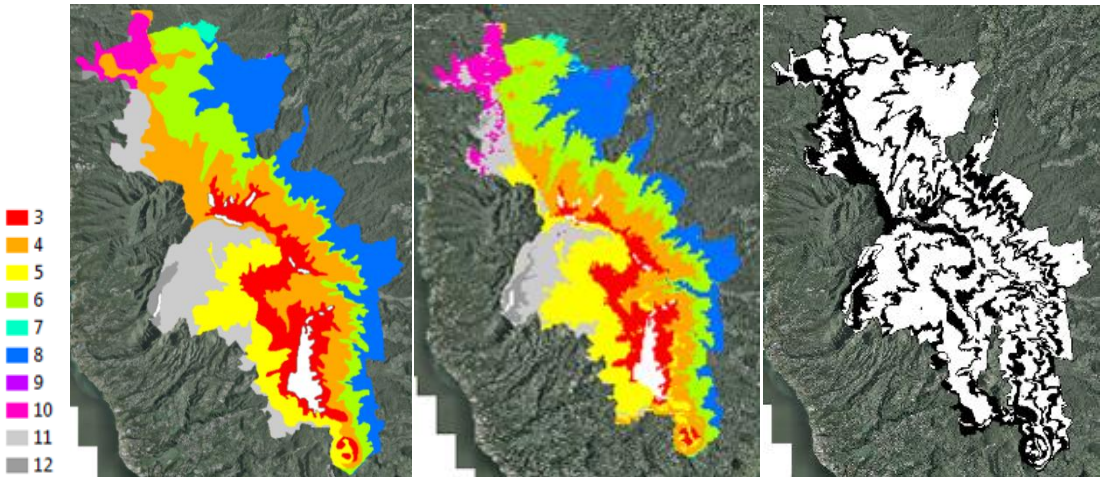
Nous avons ainsi obtenu une cartographie proche de celle obtenue en 1996. La Figure 27 présente à droite la carte écologique de référence obtenue en 1996, au milieu et à droite deux cartographies obtenues respectivement à partir d'un apprentissage par arbre de décision fonctionnel [**104**] et par arbre de décision de type C4.5 [**176**].



(a) (b) (c)

Figure 27 – (a) Carte Ecologique (1996) - Classification supervisée (b) C4.5 et (c) FT (2011)

La principale différence entre les cartographies se situe dans la zone extérieure de la Basse-Terre qui correspond à la zone côtière et agricole. La Figure 28, montre le parc naturel de la Guadeloupe classé selon la carte écologique de référence (à gauche) et la cartographie obtenue (au milieu). On peut constater une grande cohérence entre la carte de référence et les cartographies obtenues, les différentes formations étant localisées de manière correcte. La Figure 28 (à droite) montre la différence (en noir) entre les deux cartographies. Les différences sont localisées aux frontières entre les classes et ne sont pas homogènes : elles dépendent à la fois des classes et de leur localisation. Le Tableau 6 présente la matrice de confusion obtenue pour les classes représentées dans l'exemple. On constate une large domination de la diagonale qui montre la bonne cohérence de la cartographie obtenue pour l'ensemble des classes. En tenant compte du caractère manuel et arbitraire de certaines frontières de la carte de référence, nous avons pu valider la classification obtenue et l'approche proposée.



(a) *(b)* *(c)*

Figure 28 – Parc Naturel de la Guadeloupe : (a) Carte Ecologique (1996) (b) Classification supervisée FT (2011) (c) Différences (noir)

classes	3	4	5	6	7	8	10	11	12
3	**77**	10.5	12.5						
4	4.7	**71.9**	4.3	9.1			2.2	7.8	
5	1.4	2.8	**95.3**				0.2	0.3	
6		3.3		**93.7**		3			
7				19.7	**78.7**	0.7	0.9		
8				10.2	1	**88**	0.8		
10				0.2			**95.5**	3.8	0.5
11		1.1	10.4				7.6	**72.4**	8.5
12							10.8	21.2	**68**

Tableau 6 - Matrice de confusion entre *FT* et la classification de référence (% de surface)

Ces travaux ont été intégrés dans un chapitre de livre [**4**] (à paraître en décembre 2011) et une soumission dans une conférence nationale en septembre 2011 [**37**]. Ils seront également soumis à la revue *GeoInformatica* en novembre 2011.

Le même principe a été utilisé dans un autre contexte (classification d'habitats) présenté dans la section suivante en introduisant également une étape préliminaire et d'autres types de données.

Par ailleurs, cette première classification par approche vectorielle a permis de confronter la méthode de classification avec la carte de référence afin de prouver sa bonne tenue mais la classification stricte obtenue n'est pas en adéquation avec la nature transitoire des forêts. Afin d'améliorer le modèle même des forêts représentées par la cartographie nous proposons dans la section 5.3.3 un modèle vectoriel pour les données diffuses plus fiable que les modèles disponibles dans la littérature.

5.3.2.2 Classification de l'habitat dans un contexte d'épidémie de dengue

Cette étude a pour objectif d'évaluer l'impact de l'habitat en tant que vecteur du virus de la dengue et son évolution sur la période 1996-2004. L'étude est menée à l'institut pasteur de la Guadeloupe dans le cadre de la thèse de *Laurent Girdary* encadrée par *Laurence Marrama* et que je co-encadre. Mon apport dans cette étude se situe au niveau de l'utilisation des *SIG* et des méthodes de classification pour l'obtention des cartes d'habitat.

Le processus mis en place comporte deux étapes. La première consiste à définir les classes d'habitat à partir d'une classification non supervisée (analyse en composantes principales puis classification ascendante hiérarchique) réalisée sur l'ensemble des données (19 caractéristiques retenues comportant des données *INSEE* et des données géographiques). Cette étape a permis de déterminer 8 classes d'habitation (Tableau 7).

Numéro	Intitulé	Couleur
1	Touristique, PME	Bleu
2	Naturelle	Vert
3	Zone intermédiaire	Rose
4	Région agricole	Jaune
5	Zone résidentielle	Orange
6	Zone rurale	Rouge
7	Périphérie de ville	Gris
8	Centre urbain	Noir

Tableau 7 - Classes d'habitat

La seconde étape consiste à réaliser une classification supervisée des habitats et est approximativement la même que dans l'étude précédente. Nous avons donc successivement

1. fusionné des couches d'information telles que la carte des sections de la Guadeloupe et les zones classées urbaines, agricoles et naturelles.
2. mis en relation ces informations spatiales avec des données démographiques de l'*INSEE* telles que le type et la taille des habitations, les équipements de confort (climatisation, voiture, etc.), la population résidente.
3. calculé des informations géostatistiques à partir des relevés de bâtit (densité d'habitation, distance inter habitation, etc.).

4. réalisé un apprentissage sur des zones identifiées de chaque type d'habitat.
5. appliqué une classification supervisée sur l'ensemble des unités vectorielles afin de les labelliser.

Trois jeux de données sont disponibles respectivement pour les années 1996, 2000 et 2004.

Les résultats ont permis de constater que les classifications obtenues pour les différentes années sont cohérentes et que l'on retrouve la même répartition des habitats avec les deux méthodes d'apprentissage (la Figure 29 illustre les résultats obtenus).

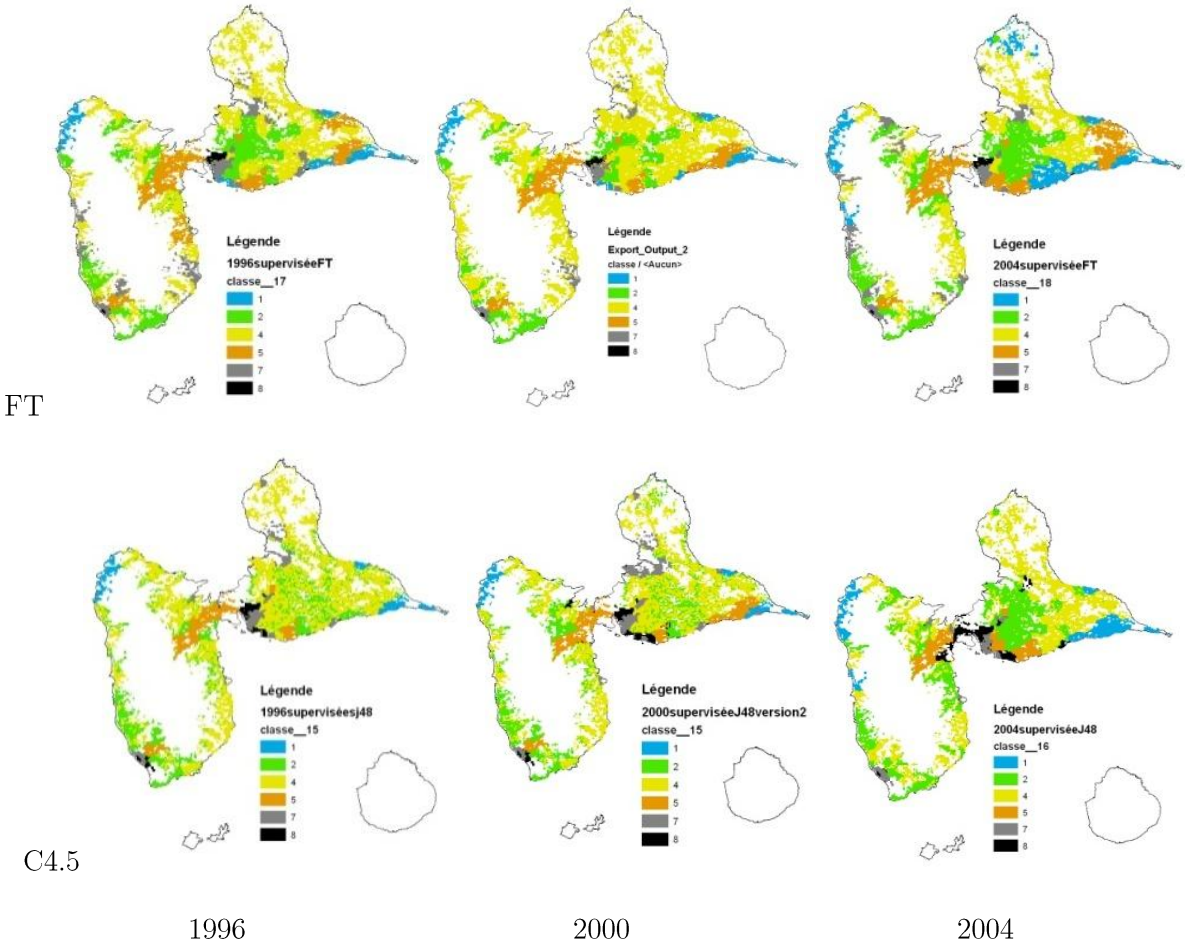


Figure 29 - Classification des habitats en 1996, 2000 et 2004 avec les méthodes d'apprentissage FT et C

Par ailleurs, ces travaux ont permis de faire ressortir des tendances à long terme dans l'évolution de l'habitat et de déterminer une méthode stable pour la classification des habitats. Ces travaux ont été soumis dans la revue *International Journal of Geographical Systems* [**38**] (septembre 2011).

5.3.3 Adaptation des formats vectoriels pour les données diffuses

Dans les conclusions du numéro spécial *Spatial Data Types for Database Systems* de *Lecture Notes in Computer Science,* en 1997 [**185**], les auteurs soulignent l'importance des structures de données dans les *SIG* et leur manque d'adaptation à certaines données (données

floues). Même si plusieurs pistes semblaient se dégager à l'époque, la difficulté d'implémentation, de modélisation et la complexité des calculs ont conduit, plus de 10 ans après à une avancée très minime sur le sujet [**108**].

Nous nous intéressons donc ici aux données dont les frontières sont naturellement diffuses et qui sont difficilement modélisables avec des ensembles stricts (seul modèle permettant de représenter des objets dans les *SIG*). Nous pouvons citer comme exemple : les phénomènes naturels (forêts, déserts, sol, géologie, séparation entre une montagne et une vallée, biotopes, pluie, inondation, vent, etc.), sociaux (changement de densité de population, cartes géopolitiques, etc.) ou encore culturels ou religieux.

Historiquement, la première approche pour traiter ces données a été de les faire restreindre à des objets stricts en réalisant des approximations ou des choix arbitraires sur les frontières (c'est le cas de l'approche présentée dans la section 5.3.2.1 sur la classification des forêts) mais l'émergence de problématiques plus fines que la simple consultation de cartes approximatives faisant intervenir de nombreux critères a mis en défaut cette première approche qui a conditionné la structuration même des données (modèle conceptuel) au sein des *SIG*.

Pour résoudre le problème, nous devons nous intéresser à plusieurs aspects des données : leur stockage, leur manipulation et leur visualisation. Plusieurs approches sont possibles : la théorie des ensembles flous, la théorie des fonctions de croyance [**64**], les espaces discriminants [**109**].

Les premiers travaux sur la modélisation floue en 2D ont été menés par *Peter Burrough* en 1986 [**71**]. Depuis les applications sur des données spatiales se sont multipliées et le contexte technologique offert par les *SIG* a permis de les développer. On a ainsi vu successivement la définition des opérations d'union et d'intersection floues, des relations spatiales floues ([**67**], [**139**]) dans un premier temps pour des données quantitatives [**113**] (degré d'appartenance) puis pour des données qualitatives ([**67**], [**182**]). Tous ces travaux ont été menés sur des données de type *raster* ([**67**], [**164**], [**182**], [**177**], [**184**], [**192**], [**193**],[**218**]). Il est donc nécessaire de rastériser les données vectorielles avant de pouvoir leur appliquer des opérateurs. Ceci implique que l'on choisisse une échelle d'analyse, que l'on rastérise les données (passage d'une représentation vectorielle à une grille), puis que l'on applique les outils d'analyse floue. Cette approche a comme inconvénients *(i)* une perte d'information lors de la rastérisation conduisant à une diminution de la précision, *(ii)* un traitement lourd si on applique cette analyse à une grande échelle, *(iii)* l'absence de représentation floue des données sources. Ce dernier point a souvent été soulevé comme inconvénient dans les Sciences de l'Information Géographique ([**58**],[**84**],[**97**][**123**], [**124**], [**186**],[**213**]).

L'enjeu actuel est donc de pouvoir traiter directement les données vectorielles afin de permettre une meilleure précision et une meilleure abstraction des données. Seules quelques rares études ont introduit cette dimension vectorielle aux géo-traitements flous ([**65**], [**126**]) et de nombreuses recherches sont encore à mener, notamment sur la modélisation des données. Ces approches consistent en une modélisation d'un ensemble flou par une série de zones tampon régulières internes ou externes à un polygone strict. Chaque zone définit une zone ayant même valeur pour la fonction d'appartenance de l'ensemble dit flou. Le principal inconvénient de cette approche est que les zones tampon sont des couronnes équidistantes des contours du polygone central ce qui ne traduit pas la réalité des phénomènes observés. En effet, dans le cas des

formations forestières par exemple, la rapidité de transition d'une formation à une autre dépend de paramètres physiques (tels que le relief) qui ne varient pas de la même manière dans toutes les directions.

Les travaux présentés dans ce cadre ne sont que les premiers d'une longue série détaillée dans les perspectives (section 6 page 82). Néanmoins, nous présentons ici le premier algorithme permettant d'obtenir un modèle vectoriel flou ainsi que les premiers résultats concernant la représentation des formations forestières. On considère un ensemble de n classes (C_k , $k = [1, n]$)
Le processus est le suivant :

1. On définit des paramètres flous caractérisant les formations forestières : pour cette étape, nous avons identifié un premier jeu des paramètres topographiques influençant les formations forestières, ils sont issus de l'étude sur la classification vectorielle (section 5.3.2.1) (altitude, pente, versant, exposition, latitude) (on note $L = \{L_1, \dots, L_N\}$, l'ensemble des N couches topographiques).
2. On partitionne l'espace en unités homogènes vectorielle ($UV = \{UV_i, i = \{1, M\}\}$). Dans le cas des formations forestières ceci correspond aux unités obtenues lors de la classification vectorielle (section 5.3.2.1). On note $A \subset$ UV un ensemble d'apprentissage et de test
3. On associe chaque UV_i à un vecteur d'attributs $F = \{F_{i_1}, F_{i_2}, \dots, F_{i_N}\}$.
4. On utilise des arbres de décision fournissant pour chaque UV_i à classer un vecteur de degrés de confiance $\mu_i = \{\mu_{i1}, \dots, \mu_{in}\}$ indiquant la confiance dans le classement de l'UV_i à chacune des classes. On peut donc représenter pour chaque classe une carte vectorielle floue basée sur les valeurs des degrés de confiance $m_k = \{\mu_{1k}, \dots, \mu_{Mk}\}$.

Ces premières cartes sont vectorielles mais bien trop morcelées pour être exploitables facilement. La Figure 30 présente les résultats obtenus pour une évaluation floue des différentes formations forestières selon le principe précédent. Les valeurs élevées de la fonction d'appartenance sont représentées en blanc et les valeurs basses en noir. On constate que même si certaines formations sont assez nettement localisées (1, 4 et 9 par exemple) il existe de vastes zones de recouvrement pour lesquels on rencontre dans une même zone géographique des valeurs élevées de la fonction d'appartenance de plusieurs formations. Ceci montre la complexité des formations et le caractère progressif des transitions d'une formation à l'autre à même d'engendrer des incertitudes sur le positionnement des frontières entre formations.

Cette première représentation peut être conservée comme représentation interne des structures floues mais la multitude de polygones constituant chacune des formations doit être simplifiée. Les paramètres définissant une vue v simplifiée d'une classe vectorielle floue sont le nombre de groupements g_v et les bornes des intervalles correspondants $I^v = \{I_0^v, I_2^v, \dots, I_{g_v}^v\}$ avec $I_0^v = 0$ et $I_{g_v}^v = 1$. On peut ainsi par exemple mettre en relief les UV ayant les coefficients de confiance les plus élevés avec plus ou moins de finesse.

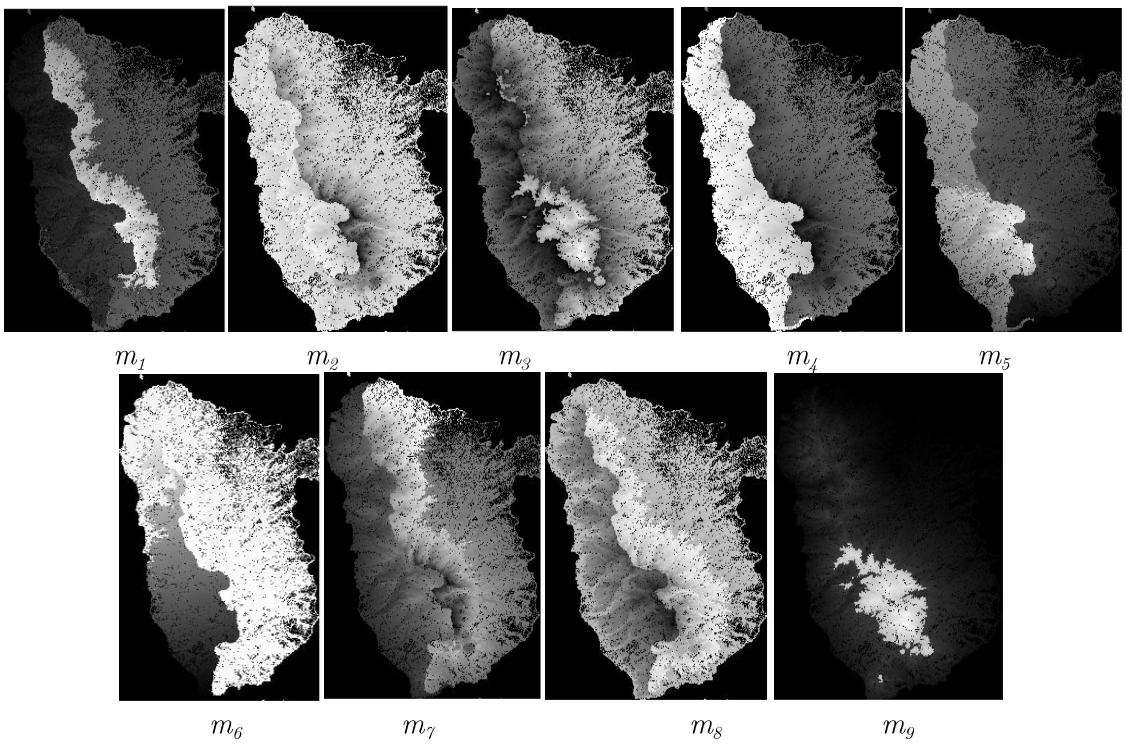


Figure 30 - Cartes Floues de 9 formations forestières

La Figure 31 donne plusieurs représentations simplifiées de la formation forestière m_9 de la Figure 30. Les trois premières font intervenir 5 zones tampons ($g_v = 5$), les 3 suivantes 10 zones ($g_v = 10$) tampon. Ceci signifie donc que l'on a découpé les valeurs de la fonction d'appartenance en autant d'intervalles et formé des ensembles vectoriels par agrégation à partir des unités vectorielles correspondantes. La définition des intervalles a une influence sur les zones tampon et doit être faite en fonction des applications visées.

Dans la colonne de gauche on trouve un découpage en intervalles de même taille. Dans la colonne centrale un découpage irrégulier mais permettant d'obtenir un nombre d'unité du même ordre de grandeur dans chaque intervalle. Dans la colonne de droite les intervalles sont également irréguliers mais concentrés sur les valeurs élevées de la fonction d'appartenance. Par exemple, la représentation au milieu de la première ligne est centrée sur les valeurs les plus nombreuses $I = \{0, 0.03, 0.07, 0.13, 0.22, 1\}$, celle de droite de la deuxième ligne est centrée sur les plus hautes valeurs $I = \{0, 0.14, 0.5, 0.75, 0.79, 0.82, 0.86, 0.91, 0.95, 0.98, 1\}$.

La colonne de droite est la plus intéressante pour ce qui concerne nos applications puisque l'on souhaite avoir une plus grande finesse d'analyse dans les zones pour lesquels on a une grande confiance en l'appartenance de celles-ci à la formation. Mais la recherche de zones intermédiaires peut nous conduire à recentrer les intervalles sur les valeurs proches de 0.5.

On note bien, à la différence des zones tampons utilisées dans la littérature, que les zones définies sont non régulières et représentent les données de manière plus fiable (meilleures représentation des valeurs réelles du degré de confiance).

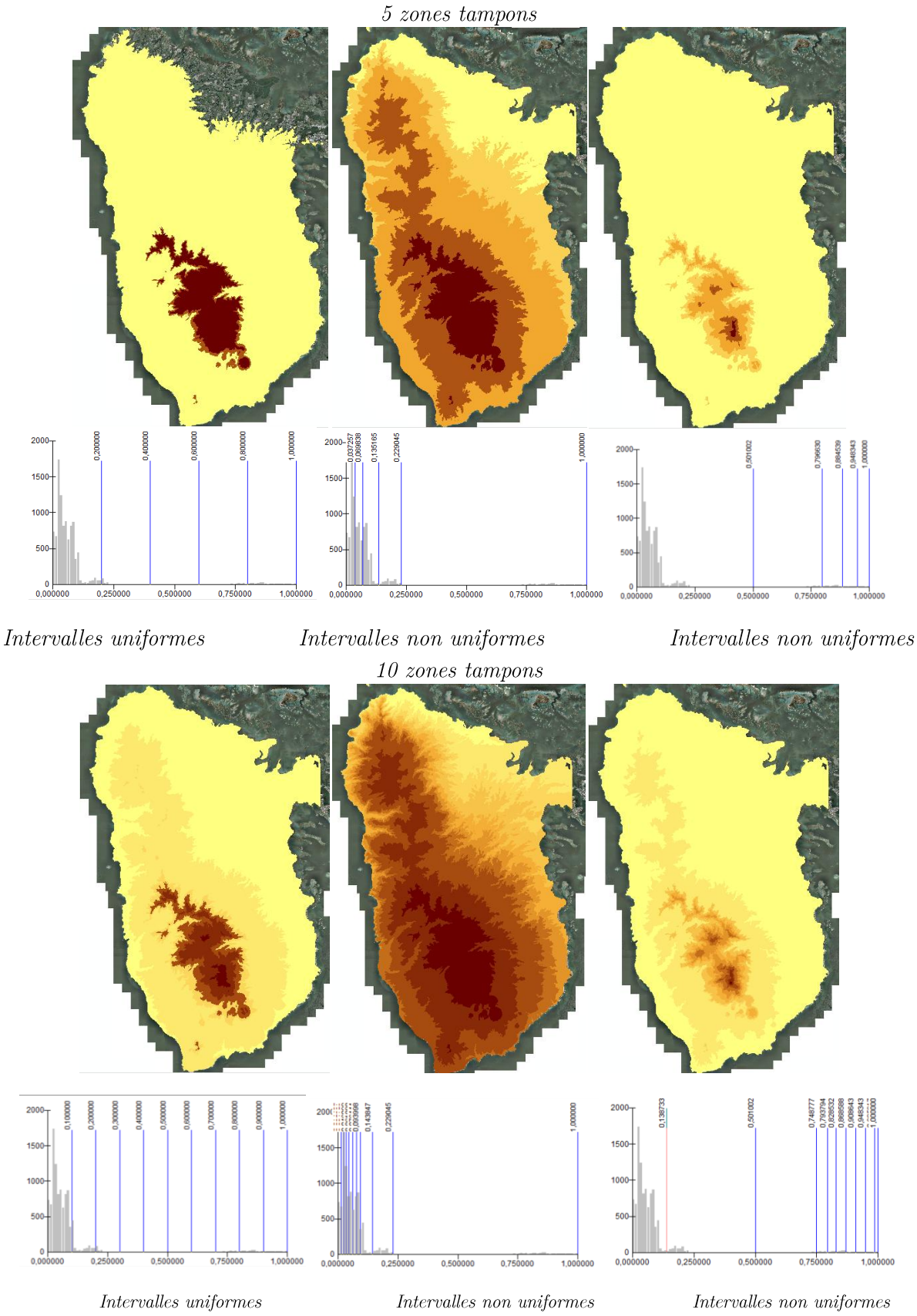


Figure 31 - Différentes représentations vectorielles simplifiées d'une formation forestière (classe 9)

Les valeurs des coefficients de confiance présentées dans la Figure 30 montrent que certaines formations ne sont pas bien localisées ou du moins possèdent des zones de transition très étendues avec d'autres formations. Nous nous intéressons maintenant à la modélisation de ces zones de transition de manière à en déduire, si nécessaire, des classes de transition. Ces zones sont localisées en se basant sur un seuillage des degrés de confiance de la manière suivante :

1. On fixe un seuil S
2. Pour chaque UV_i on trie les classes par ordre décroissant de degrés de confiance : $\{\mu_{ij_1} \geq \mu_{ij_2} \geq \cdots \mu_{ij_n}\} \rightarrow \{C_{j_1}^i \ , C_{j_2}^i \ , \ldots, C_{j_n}^i \ \}$
3. On attribue à chaque UV_i la liste minimale des classes définie par $E_i(S) = \{C_{j_1}^i \ , C_{j_2}^i \ , \ldots, C_{j_p}^i \ | \ \sum_{k=1}^{p} \mu_{ij_k} > S\}$

Pour tout UV_i , $E_i(0) = \{C_{j_1}^i\}$, nous réobtenons dans ce cas une classe unique pour chaque UV ce qui revient à produire une classification stricte à partir d'un ensemble de n classes floues en choisissant la classe ayant le degré de confiance le plus élevé. Cette classification correspond à la classification obtenue dans la section 5.3.2.1. De manière générale, si $E_i(S)$ est un singleton, c'est qu'il n'y a pas d'ambiguïté sur la classe. Dans le cas contraire, il s'agit d'une zone de transition. En fonction de la valeur du seuil, les zones de transition sont plus ou moins nombreuses et étendues mais elles ne sont ni constantes, ni régulières dans l'espace.

En faisant varier S de *0* à *1* nous pouvons hiérarchiser l'ordre d'apparition des transitions. Plus une transition apparait pour une valeur faible de S, plus elle témoigne de l'incertitude sur l'affectation des classes dans la zone concernée. Par ailleurs, plus une transition est étendue et plus la transition entre les classes concernées est lente. Le choix d'une valeur de S permet de figer les transitions à prendre en compte.

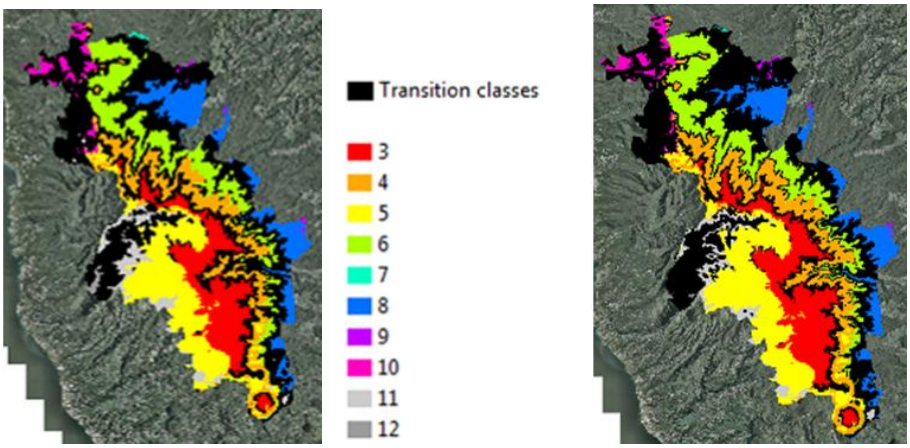
Théoriquement, toutes les transitions entre classes sont possibles et la hiérarchie pourrait faire apparaître jusqu'à 2^n transitions mais dans notre cas, la cohérence spatiale des classes que nous étudions limite le nombre de transitions possibles. Cette cohérence est au fait que les données spatiales ne sont pas indépendantes et stationnaires [**59**] et la définition des modèles dépend précisément de la localisation.

Le Tableau 8 présente la hiérarchie des transitions obtenues en faisant varier S de *0* à *1* dans le cas de la classification des forêts. Pour chaque valeur de S l'ensemble des transitions est indiqué puisque certaines peuvent apparaitre et d'autres être englobées dans des transitions plus importante (par exemple la transition *10-12* présente pour $S = 0.9$ a été remplacée par la transition *10-11-12* pour $S = 1$). Dans cet exemple il y a entre *3* et *13* classes de transitions. Les premières transitions à apparaitre sont celles entre les classes *6, 7* et *8*. La transition *6-8* s'étend également rapidement avec l'augmentation de S *(*colonne de droite du tableau). C'est également le cas de la transition *10-11* qui apparait pour $S = 0.5$ et qui s'étend rapidement. Etant donnée leur apparition pour des valeurs faibles de S et leur étendue, ces classes de transition présentent un intérêt puisqu'elles font apparaitre de vastes zones d'incertitude.

S	Transitions	Superficie
0	ϕ	0
0.1	**6-7, 6-8, 7-8**	1.2 %
0.3	6-7, 6-8, 7-8**, 10-12, 11-12**	2.1 %
0.5	6-7, 6-8, **6-11**, 7-8, **8-11, 10-11,** 10-12, 11-12	4.4 %
0.7	**3-4, 4-6,** 6-7, 6-8, 6-11, 7-8, 8-11, 10-11, 10-12, 11-12	21.4 %
0.9	3-4, 4-6, **4-11**, 6-7, 6-8, 6-11, 7-8, 8-11, 10-11, 10-12, 11-12	25.1 %
1	3-4, 4-6, 4-11, 6-7, 6-8, 6-11, 7-8, 8-11,10-11, 11-12, **6-7-8, 6-8-11, 10-11-12**,	27.2 %

Tableau 8 - Hiérarchie des transitions

La Figure 32 représente les classes de transition (en noir) pour deux valeurs de seuil S. On constate que certaines transitions sont très étendues et d'autres très localisées. Fixer une frontière dans une zone bien localisée peut avoir un sens (et est souvent rattaché à un changement brusque de topologie comme une ligne de crête) mais dans des zones étendues (transitions entre les classe 6 et 8 par exemple) le choix ne pourra être qu'arbitraire.



S=0.8 *S=0.9*

Figure 32 - *Classes de transition*

Nous préconisons, dans le cas de zones de transition étendues, la création de classes de transition entre les classes existantes. Ces classes de transition traduisent plus qu'une incertitude sur une classification puisqu'elles représentent véritablement le modèle de passage d'une formation forestière à une autre qui suit un gradient plus ou moins rapide. Les critères permettant de décider de la création ou non d'une classe de transitions sont présentés dans les perspectives.

Par ailleurs, les classes présentes sont rattachées à un concept dans l'ontologie présentée à la section 5.3.1 page 58. Une hiérarchie existe donc entre ces classes et sous couvert du respect de cette hiérarchie imposée par l'ontologie, les classes de transition peuvent être nommées par le concept généralisant ceux associés aux classes.

Une partie de ces travaux ont été soumis dans [**4**] et [**37**]. Une autre soumission plus importante est envisagée dans la revue *GeoInformatica* prochainement.

5.3.4 Modélisation sémantique des scènes

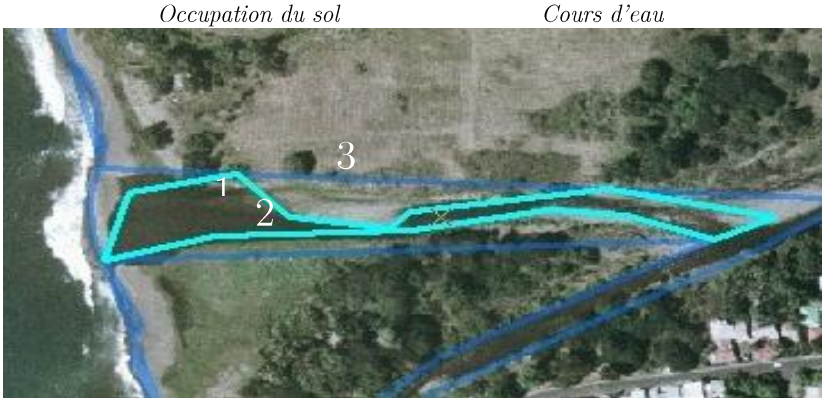
La dernière contribution présentée dans le cadre de la modélisation de données concerne le problème de mise en correspondance de différentes représentations d'une même entité plusieurs fois évoqué au cours de ce mémoire. Le principal obstacle pour ces mises en correspondance provient des différences de modélisation des entités et des incertitudes quant à leur localisation. *Goodchild* indiquait en 2010 [**108**], lors de sa rétrospective des avancées des *SIG* durant les 20 dernières années, que parmi les défis à relever dans les 10 prochaines années, l'intégration du caractère incertain des frontières, de leur sémantique et l'organisation des bases de données en conséquence était des enjeux majeurs.

Nous avons défini dans ce cadre la notion de couche sémantique qui permet de décrire un modèle abstrait d'une scène. Ce modèle de couche a été proposé dans [**10**] et permet de lever les ambigüités lors d'un conflit en choisissant la représentation la plus adaptée. La Figure 33 illustre ce principe avec deux couches thématiques, l'une représentant l'occupation du sol et l'autre représentant les cours d'eau sous forme de polygone. Les zones *agricole*, *naturelle* et *urbaine* extraites de la couche d'occupation du sol sont délimitées en partie par une rivière mais les frontières correspondantes ne coïncident pas avec les limites issues de la couche des cours d'eau.

Dans un premier temps, l'introduction d'une sémantique au sein des couches permet de rattacher chacune des frontières (ou partie de frontière) à un concept dans une ontologie. Lors d'un conflit, on peut alors choisir la représentation provenant de la couche thématique sémantiquement la plus proche du concept en question.



Occupation du sol *Cours d'eau*



Conflit

Figure 33 - Différentes représentations d'une même entité

Dans l'exemple de la Figure 33, la frontière sud de la zone agricole est rattachée au concept *rivière,* lui-même rattaché au concept *ressources en eau*, la couche représentant les *cours d'eau* est également rattachée à ce même concept *ressources en eau* alors que la couche *occupation du sol* est rattachée aux concepts *activité humaine et végétation* (Figure 34). Lors du conflit c'est donc la représentation provenant de la couche thématique sur les cours d'eau qui sera jugée la plus pertinente, puisque sémantiquement plus proche du concept rattaché à l'entité en conflit.

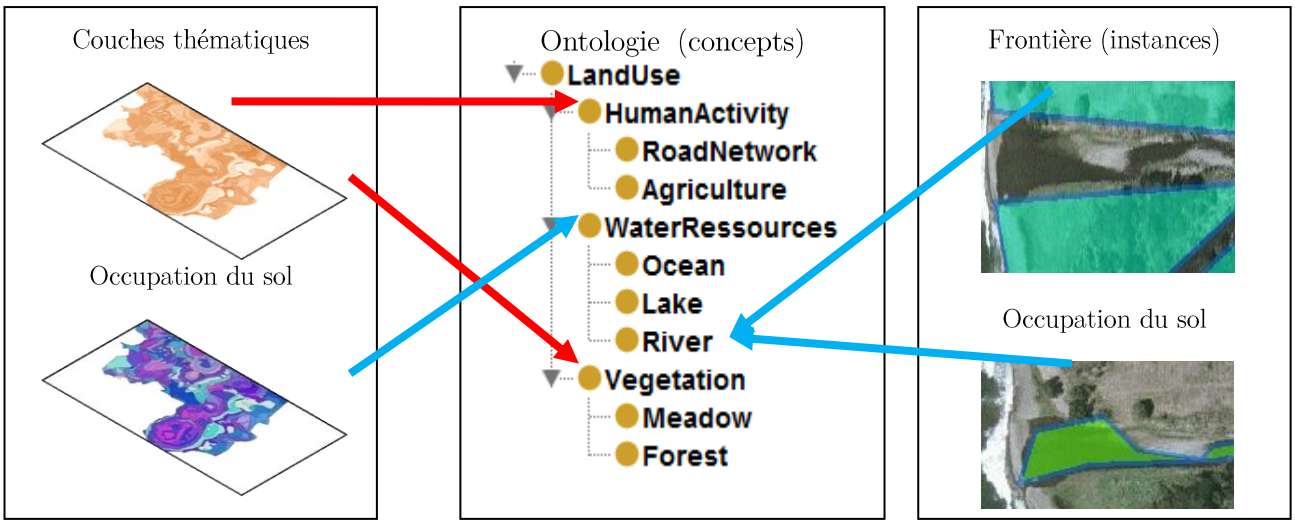
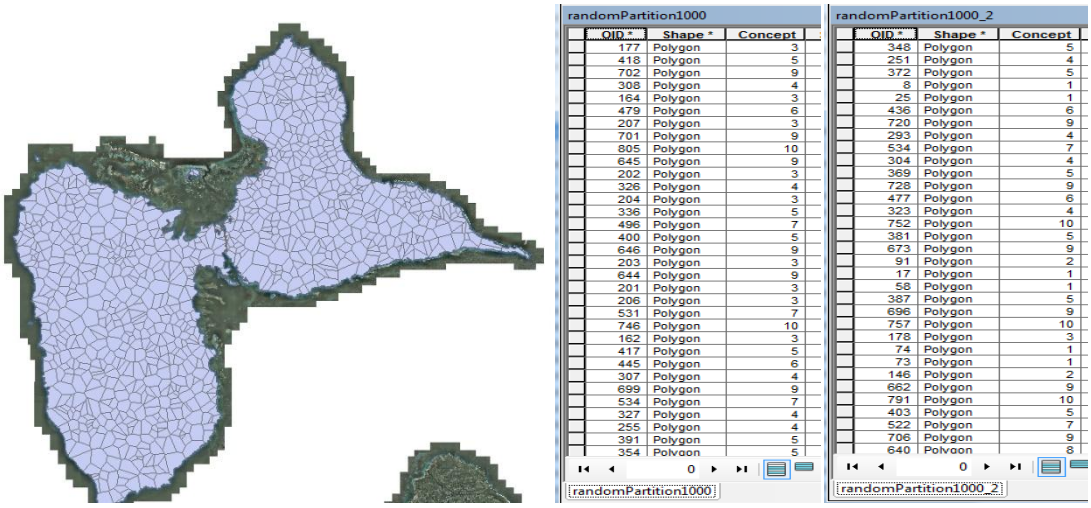


Figure 34 - Liaison couche thématique et ontologie

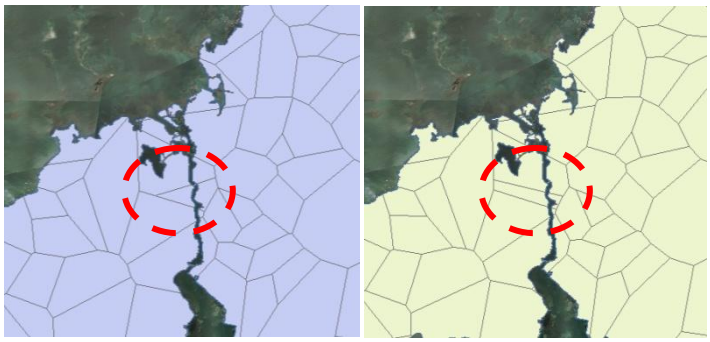
La mise en place d'un tel procédé requiert l'ajout d'une référence pour chaque entité vers un concept d'une ontologie partagée par les différentes couches thématiques indiquant la nature de l'entité (objet ou frontière). Sous réserve que les différentes ontologies partagent les mêmes termes, plus le concept est précis, plus les conflits pourront être résolus facilement. Par exemple, un lien vers un concept abstrait *cours d'eau* pourra générer une ambiguïté lors de la fusion alors qu'un lien vers une instance d'un *cours d'eau* (*rivière moustique* par exemple) permettra de lever toute ambiguïté.

Des tests ont été réalisés sur des données simulées de la manière suivante (Figure 35). On génère aléatoirement un ensemble de polygones de manière à partitionner la *Guadeloupe* $P_1 = \{P_i^1 \mid i = [1,n]\}$, chaque polygone est composé d'un certain nombre de frontières partagées avec d'autres polygones ($P_i^1 = \{F_j^i \mid j = [1,N_i]\}$ (a). On génère également aléatoirement deux arbres de concepts ($A_k = \{C_l^k \mid l = [0,m]\},\ k = \{1,2\}$) et on associe aléatoirement un concept (en provenance de l'un des deux arbres) à chacune des frontières ($\{F_j^i,\ C_{ij}\}$ avec $C_{ij} = C_l^k \mid l \in [1,m],\ k \in \{1,2\}$)(b). On génère ensuite une perturbation δ_{ij} sur la position des frontières pour créer une deuxième partition ($P_2 = \{P_i^2 \mid i = [1,n]\}$ *avec* $P_i^2 = \{F_j^i + \delta_{ij} \mid j = [1,N_i]\}$) de la Guadeloupe les concepts n'étant pas perturbés ($\{F_j^i + \delta_{ij},\ C_{ij}\}$) (d). On rattache ensuite chacune des partitions créée $\{P_1, P_2\}$ au concept racine d'un des arbres $(\{P_k, C_0^k\} \mid k = \{1,2\})$. Parmi les frontières présentes dans une partition, certaines sont donc rattachées à un concept dérivé de celui rattaché à la partition elle même ($\{F_j^i \mid i \in [1,n], j \in [1,N_i],\ C_{ij} = C_l^1,\ l \in [0,m]\}$ pour la partition P_1et $\{F_j^i + \delta_{ij} \mid i \in [1,n], j \in [1,N_i],\ C_{ij} = C_l^2,\ l \in [0,m]\}$ pour la partition P_2)

et d'autres à un concept de l'arbre associé à l'autre partition ($\{F_j^i \mid i \in [1,n],\ j \in [1,N_i],\ C_{ij} = C_l^2,\ l \in [0,m]\}$ pour la partition P_1 et $\{F_j^i + \delta_{ij} \mid i \in [1,n], j \in [1,N_i],\ C_{ij} = C_l^1,\ l \in [0,m]\}$ pour la partition P_2). Les conflits sont matérialisés par des frontières différentes pour des concepts identiques. Etant donnée la construction des partitions, dès lors que $\delta_{ij} \neq 0$, il y aura un conflit (Figure 35 (c) et (d)).



(a) Génération aléatoire de polygones *(b) Génération aléatoire de concepts*



(c) Polygones non perturbés *(d) Polygones perturbés*

randomPartition1000_2_sel_line

QID *	Shape *	FID_randomPartition1000_2	Concept	Classe	FID_randomPartition1000_23	Concept	Classe	Shape_Length
2799	Polyline	121	2	2	121	2	2	0,334517
2800	Polyline	121	2	2	378	5	5	0,334517
2801	Polyline	378	5	5	121	2	2	0,334517
2802	Polyline	378	5	5	378	5	5	0,334517

(e) Rattachement de chaque segment d'un polygone à un concept

Figure 35 - Résolution de conflits sur des données simulées

Les tests ont été réalisés à partir d'une génération de *1000* polygones et *10* concepts différents par arbre. Lorsque l'on réalise la fusion des deux couches (polygones perturbés et polygones non perturbés) on obtient en moyenne *7543* polygones dans la couche résultante alors que le résultat attendu est de *1000* polygones (les deux couches représentant les mêmes entités). En appliquant le principe de choix de la frontière la plus adaptée, la couche fusionnée possède exactement *1000* polygones. Les résultats prouvent le bon fonctionnement de la méthode mais ne

peuvent être extrapolés dans le cas de données réelles. En effet, les données simulées permettent de résoudre tous les conflits puisque chaque frontière est renseignée et possède une représentation idéale. Les taux de résolution des conflits dépendront des annotations des frontières par des concepts. Des pistes pour générer des concepts à partir des instances d'une couche sont présentées dans les perspectives (section 6) ainsi que l'intégration des relations spatiales.

5.4 Exploitation des données

Nous présentons maintenant la dernière partie de la chaîne de traitement avec les contributions concernant l'exploitation de l'information. Ces contributions viennent en bout de chaîne lorsque les différentes sources de données sont formatées (informations spatiale et sémantique réunies sous un format vectoriel) et éventuellement corrigées (mise en correspondance des différentes représentations d'une même entité).

La première partie des travaux s'intéresse au problème de la sélection optimale de l'information lors du croisement d'une multitude de couches. La seconde partie s'intéresse à l'utilisation conjointe de l'information raster et vecteur pour produire de nouvelles informations.

5.4.1 Optimisation de la sélection d'information

Ces travaux ont commencé en 2008 avec le stage de Master 2$^{\text{ème}}$ année de *Franck Duhamel* et sont poursuivis dans le cadre de la thèse de *Wilfried Segretier* depuis 2010. Ils ont donné lieu à deux publications ([**12**] et [**14**]) ainsi que deux soumissions ([**8**] et [**41**]). Par ailleurs, les techniques d'optimisation utilisées ont été expérimentées durant ma thèse et mes premières années en tant que Maître de Conférences et ont donné lieu à 4 publications ([**6**], [**23**], [**24**] et [**25**]) autres que mon manuscrit de thèse [**35**].

L'intérêt pour la sélection d'information vient d'un problème posé par le *Parc National de la Guadeloupe* en 2008. Le problème posé était de simplifier l'information d'un ensemble d'*Unités Ecologiques Vectorielles* (*UEV*) afin de choisir les emplacements des placettes de surveillance. Les travaux actuellement réalisés dans ce cadre sont : la modélisation du problème, la définition des algorithmes de résolution et des critères ainsi que leur codage, des tests sur des données simulées.

Les techniques utilisées sont généralisables dans le cas de l'extraction ou de la découverte de connaissance (*KDD*). Dans notre contexte, lors du croisement de nombreuses couches d'information, l'information résultante est souvent difficile à exploiter car trop morcelée. Outre les erreurs de croisement liées aux incertitudes, la combinaison des couches conduit à des polygones de trop petite taille pour qu'une carte soit visuellement représentative et exploitable lorsque l'information représentée est trop riche.

Nous avons donc mis en place un algorithme d'optimisation permettant de regrouper certaines entités en tenant compte des informations sémantiques et spatiales.

5.4.1.1 Modélisation du problème

L'information contenue dans une couche est structurée à l'aide d'une ontologie. C'est-à-dire une hiérarchie de concepts tels que détaillé dans le formalisme que nous avons proposé dans [**8**] et [**41**].

Le problème est donc posé de la manière suivante. Soit $E = \{(L_1, O_1), (L_2, O_2) \dots, (L_n, O_n)\}$ un ensemble de n couples tels que L_i soit une couche d'information (Layer) et O_i une ontologie associée. Une ontologie est un ensemble de concepts structuré sous forme d'arbres $Oi = (\{C_{i,1}, \dots, C_{i,n_i}\}, A_i)$ où n_i désigne le nombre de concepts de l'ontologie $O_i, C_{i,j}$ le concept j de l'ontologie O_i *et* A_i l'arbre des concepts. Les concepts $C_{i,j}$ sont les concepts feuille de l'arbre (les concepts les plus spécifiques). Chaque nœud de l'arbre (qui peut également être considéré comme un concept) représente une généralisation des nœuds fils. Le nœud racine est donc le concept le plus général.

Il faut noter que plusieurs ontologies sont possibles pour une même couche d'information (plusieurs ensembles de concepts et plusieurs arbres). La Figure 36 représente deux arbres possibles pour un même ensemble de concepts. L'arbre de gauche présente plusieurs généralisations possibles pour un même niveau dans l'arbre alors que l'arbre de droite n'en autorise qu'une seule.

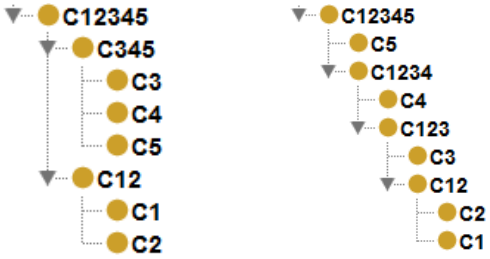


Figure 36 - Différents arbres pour un même ensemble de concepts

Le choix d'une ontologie aura un impact important sur le résultat du processus de sélection de l'information, puisque l'ontologie conditionne les regroupements. L'arbre de droite permet une plus grande finesse dans les regroupements puisqu'il permet d'envisager 1, 2, 3, 4 ou 5 concepts pour la couche, alors que l'arbre de gauche ne permet d'envisager que 1, 2 ou 5 concepts. On peut passer artificiellement de l'arbre de gauche à l'arbre de droite en ordonnant les regroupements en se basant sur une distance sémantique entre les concepts feuille ou sur d'autres attributs des concepts (comme par exemple la distance moyenne entre les instances des concepts pris deux à deux).

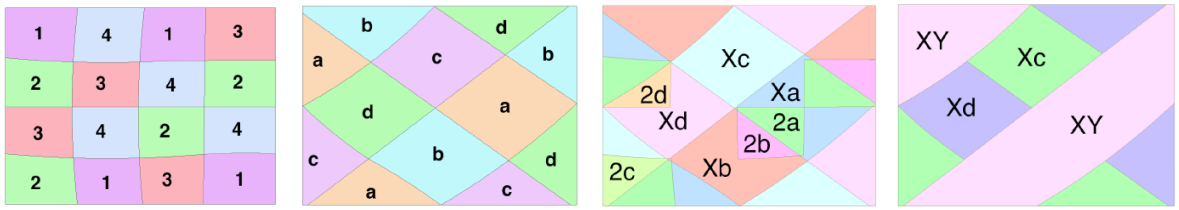
Chaque entité (instance) présente dans une couche est rattachée à un concept feuille de l'ontologie. Chaque nœud de l'arbre possède donc des informations telles que le nombre de nœuds fils, le nombre d'instances rattachées, des informations brutes sur les instances (forme et positionnement spatial) et d'autres informations calculables (surface et surface moyenne, adjacence, distance spatiale et sémantique entre les concepts ou instances). L'utilisation d'une ontologie sur les objets présents permet de garantir la cohérence de l'information, l'intégration de la spatialisation des données permet de garantir l'amélioration de sa représentation.

Pour simplifier l'information, nous devons donc tenir compte de plusieurs contraintes :

1. D'une part la qualité de l'information doit être conservée. On entend par là, une information représentative et comportant un maximum de concepts.
2. D'autre part, l'information doit être visuellement exploitable ce qui nécessite d'obtenir des entités (polygones) qui ne soient pas trop morcelées et dispersées. On doit donc tenir compte dans ce cas de la répartition spatiale de l'information.

Une première modélisation de la solution au problème de la sélection d'informations est donc un ensemble $S = \{N_1, N_2, .., N_n\}$ de n entiers représentant le niveau de généralisation conservé dans chacune des ontologies : on parle d'agrégation hiérarchique [**8**]. Lorsque l'on généralise des concepts, l'ensemble des instances des concepts généralisés sont fusionnés ce qui conduit à des ensembles de plus grande taille plus facilement interprétables. Nous verrons dans les perspectives qu'une modélisation plus complexe de la solution permettrait d'obtenir des solutions plus fines.

La Figure 37 illustre le principe d'agrégation hiérarchique à partir de deux couches *(a)* et *(b)* comportant chacune 4 concepts notés respectivement {1, 2, 3, 4} et {a, b, c, d}.



(a) Couche1 *(b) Couche 2* *(c) Fusion 1 : X={1,3,4}* *(d) Fusion 2 :X={1,2,3,4}, Y={a,b}*

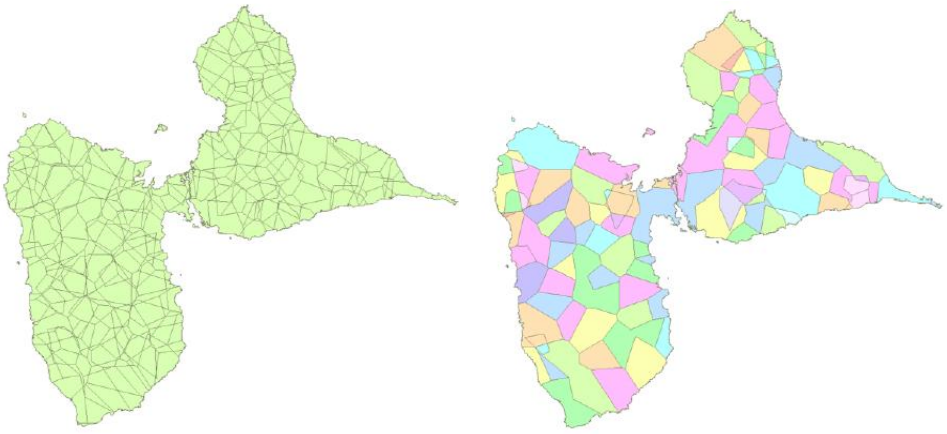
Figure 37 - Agrégation hiérarchique

5.4.1.2 Résolution du problème

La résolution du problème fait appel à des techniques d'optimisation combinatoire approchées (la recherche exhaustive n'étant pas envisageable) à population telles que les *Algorithmes Génétiques* (AG) ou à voisinage telles que la *recherche tabou* ou *le recuit simulé*. Les deux approches actuellement retenues sont les *AG* et une recherche par voisinage.

Dans le cas de l'approche par *AG*, le codage d'une solution correspond au modèle présenté précédemment. On se place dans un contexte multi-objectifs avec comme critères la maximisation du nombre de concepts conservés et la maximisation de la surface moyenne des polygones après fusion. Ces deux critères sont antinomiques puisque plus on conserve de concepts, plus l'information sera représentative mais également plus elle sera morcelée.

La Figure 38 présente des résultats obtenus sur des données simulées réalistes faisant intervenir 3 couches d'information et 30 concepts [**41**]. La solution retournée comporte 14 concepts au lieu de 30.



(a) Solution initiale *(b) solution finale*

Figure 38 - Résultat de la fusion d'information

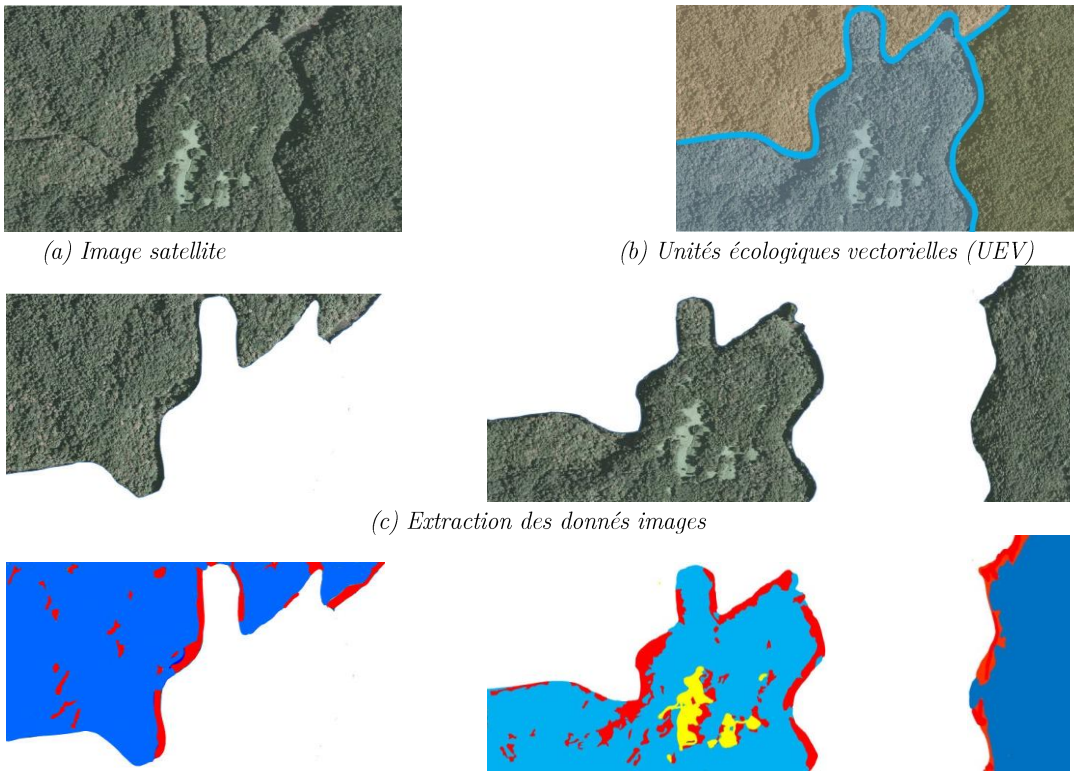
5.4.2 Coopération raster-vecteur

La dernière contribution présentée concerne la *coopération raster-vecteur*. Elle est en bout de chaîne et pourtant reboucle vers les premières contributions présentées dans la première partie. En effet, elle utilise les outils présentés pour la classification d'images satellites avec notamment la construction d'un vecteur de descripteurs faisant intervenir des informations spectrales et structurelles des images. Mais cette fois, les méthodes sont appliquées dans un cadre bien spécifique afin de tenir compte des limitations évoquées dans les conclusions de la première partie à savoir la difficulté d'appliquer un algorithme de classification sur une scène comportant un trop grand nombre de classes. On reboucle donc sur les travaux qui ont initié mon parcours de recherche pour montrer l'intérêt de la combinaison d'informations issues des images avec des informations issues d'autres sources. Ici nous utilisons les images satellites filtrées par les *Unités Ecologiques Vectorielles* (*UEV*) produites par la combinaison de couches d'information vectorielles pour analyser l'homogénéité des *UEV* et éventuellement diviser celles-ci pour produire des sous-UEV.

En effet, lorsque les *UEV* ont été obtenues à partir des couches d'information vectorielles et qu'il n'est plus possible de les diviser, elles représentent les plus petits ensembles vectoriels que l'on puisse obtenir à partir des couches d'information, elles sont donc indivisibles en utilisant uniquement ce type d'information. Néanmoins, certaines *UEV* peuvent être composées de plusieurs sous ensembles de nature différentes : forêt, glissement de terrain, habitation, etc. Si les couches d'information vectorielles n'ont pas permis la séparation de ces unités, nous pouvons utiliser les images satellites ou aéroportées afin de distinguer les différentes composantes de celles-ci.

La Figure 39 présente une image satellite (a), mise en correspondance avec des *UEV* obtenues par la fusion de plusieurs couches d'information (pente, sol, végétation) (b). Une extraction par masque a alors été appliquée sur l'image satellite afin d'obtenir les portions d'image correspondant à chaque *UEV* (c). Une classification semi-supervisée a alors été appliquée sur les sous-images de manière indépendante. Dans l'exemple de la Figure 39, les unités présentées à gauche et à droite ont été classées en deux classes : l'une représentant l'écosystème (en bleu), l'autre l'ombre (en rouge). L'unité au centre a été classée en trois classes : l'écosystème majoritaire (en bleu), l'ombre (en rouge) et un écosystème secondaire (en jaune).

Cette première étape a donc consisté à utiliser les *UEV* (information vectorielle) pour limiter l'étendue de la classification raster. La seconde étape consiste à utiliser le résultat de la classification raster pour définir une nouvelle *UEV*. La Figure 40 présente le résultat d'une vectorisation de la classification obtenue pour l'*UEV* centrale de la Figure 39. Les étapes sont les suivantes, à partir de la classification (a), on extrait une *Unité Ecologique Raster* (*UER*) (b) que l'on vectorise afin d'obtenir une *UEV* (c). A partir de cette *sous-UEV*, l'*UEV* initiale est modifiée (on supprime la *sous-UEV* par une opération de différence ensembliste) afin d'obtenir quatre *UEV* distinctes (d).



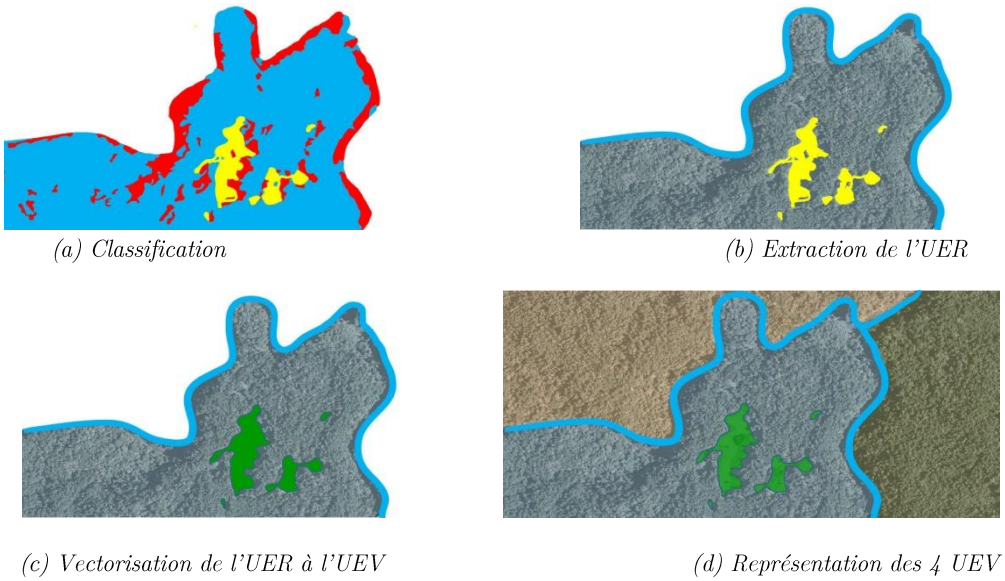
(a) Image satellite

(b) Unités écologiques vectorielles (UEV)

(c) Extraction des donnés images

(d) Classifications

Figure 39 - Classification raster des unités écologiques vectorielles



(a) Classification

(b) Extraction de l'UER

(c) Vectorisation de l'UER à l'UEV

(d) Représentation des 4 UEV

Figure 40 - Extraction d'unités écologiques vectorielles à partir d'une classification raster

L'étape suivante consiste à renseigner la nouvelle *UEV* créée avec des informations sémantiques fournies par des experts. Ces travaux ont donné lieu à une publication [**11**] et sont toujours en cours.

5.5 Synthèse de l'organisation des thématiques

Le schéma de la Figure 41 résume l'ensemble des contributions présentées dans la section 5. Les trois parties *extraction*, *modélisation* et *exploitation de l'information* constituent un ensemble d'approches et de méthodes cohérentes et complémentaires au sein d'une chaîne de traitement permettant de passer d'une *donnée raster brute* à une *information vectorielle exploitable*. D'un point de vue quantitatif, la majeure partie des contributions se situe dans la première partie, *l'extraction d'information*, ceci principalement pour des raisons historiques liées à mon parcours et pragmatiques liées à une plus grande facilité (toute relative) d'obtenir des résultats en limitant les entrées et les interactions pouvant exister entre elles dans les différentes approches explorées.

Cependant, la plus value de ces approches, à caractère très technique, peut être grandement améliorée en les plongeant dans un contexte plus formel où la sémantique de l'information joue un rôle important tant du point de vue de la modélisation de l'information que du point de vue de la mise en œuvre des méthodes. A l'issue de ces constatations et de l'existence d'un certain nombre de verrous technologiques concernant la représentation de données floues, les contributions se sont donc tournées vers la modélisation des données (seconde partie de la chaîne) avec notamment la définition d'un dictionnaire, d'une couche sémantique et d'un premier modèle vectoriel flou.

Mais c'est également pour ces différentes raisons que les dernières contributions sur ces méthodes ont porté sur l'intégration d'une sémantique dans le processus même d'extraction (structuration à l'aide d'ontologies, modélisation des connaissances expert, etc.). De la même manière, dans la partie *exploitation de l'information* l'algorithme de sélection d'information utilise la sémantique comme l'un des critères de sélection.

La seconde partie de la chaîne de traitement (*la modélisation de l'information*) est donc destinée à être développée et à être progressivement intégrée aux deux autres (ou à progressivement intégrer les deux autres ?) de manière à la rendre transparente. Mais avant d'en arriver à ce stade, et même si les premières contributions dans ce domaine sont prometteuses, il reste de nombreux verrous technologiques et de nombreuses pistes à explorer.

Après une brève présentation des perspectives à court et moyen terme sur les parties *extraction et exploitation de l'information*, le reste de ce document (section 6) s'attache à présenter certaines des pistes possibles pour réaliser l'intégration de la sémantique et des modèles dans la gestion et la représentation des données géographiques.

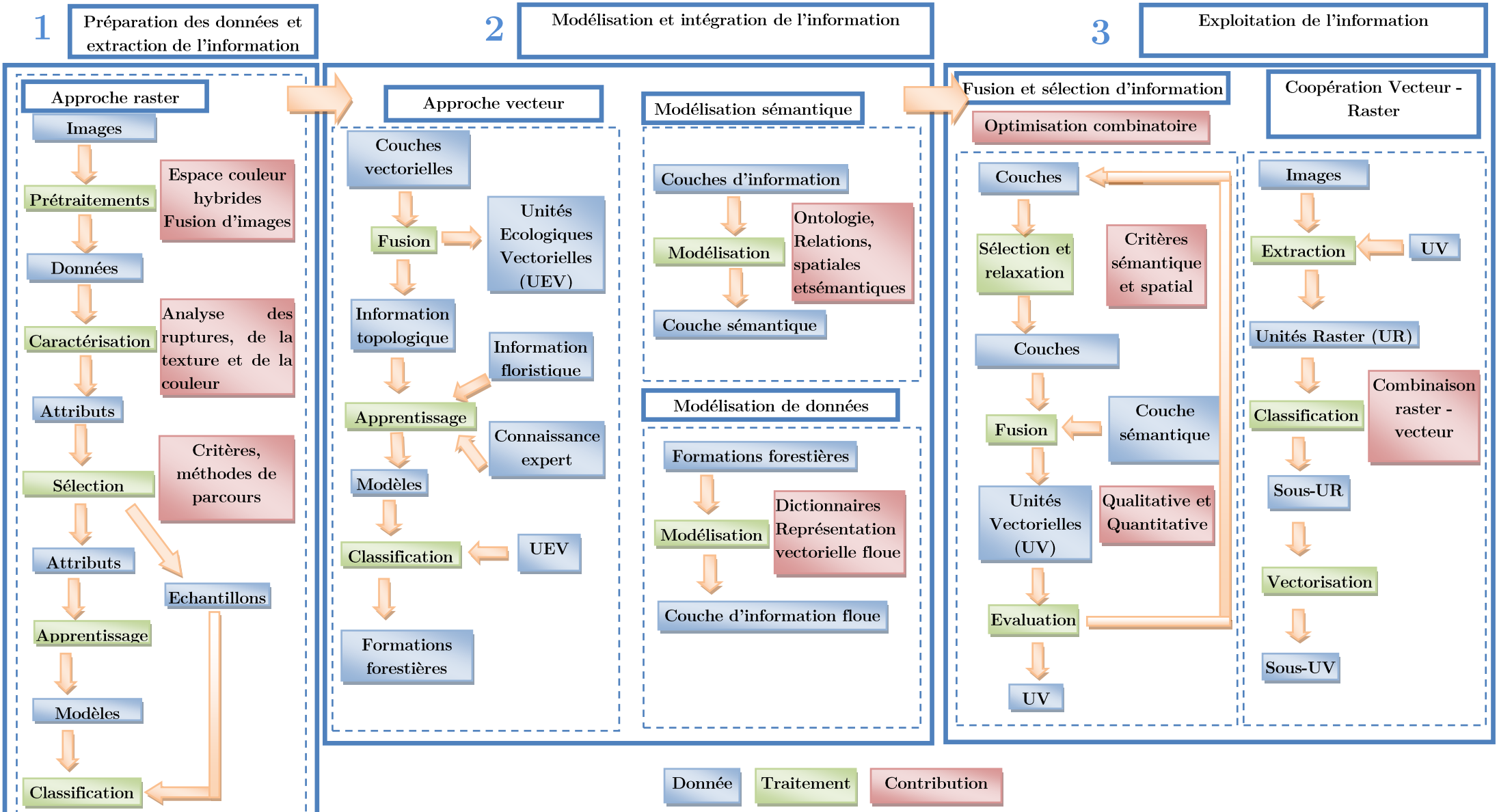


Figure 41 - Schéma synthétique complet de la chaîne de traitement

6 Perspectives

L'orientation prise depuis 2008 se confirme ici à travers les perspectives qui sont essentiellement orientées vers la modélisation des données au sein des *SIG*. Cette section est organisée de manière « chronologique ». Après une brève présentation du contexte dans lequel vont s'insérer ces perspectives (6.1) nous présentons les perspectives à court et moyen termes (6.2) sur des aspects techniques tels que la sélection d'information (6.2.1.) ou la classification vectorielle (6.2.2 et 6.2.3).

Puis nous détaillons l'axe principal sur la *modélisation des données* qui sera développé sur le long terme (6.3). L'idée transversale de toutes ces perspectives est de proposer une meilleure représentation de la réalité sous un format adapté (vectoriel) afin de faciliter la manipulation et le traitement de l'information. Nous présentons dans ce cadre des réflexions sur la représentation des données (6.3.1), sur l'intégration de la sémantique (6.3.2) et sur le formalisme associé (6.3.3). Nous présentons également des pistes pour la prise en comptes des erreurs dans les modèles (6.3.1.1), une extension du modèle vectoriel flou (6.3.1.2) ainsi que les perspectives envisagées pour l'intégration des relations spatiales (6.3.4.1) et sémantiques (6.3.4.2).

6.1 Vers une cohésion autour des *SIG*

La cohésion autour des axes de recherche présentés dans la première partie de ce mémoire se retrouve à plusieurs niveaux.

D'une part les applications de mes recherches, la protection et la valorisation des ressources forestières, sont en adéquation avec le premier axe stratégique inscrit dans le *Plan d'Action Stratégique* de l'UAG *Développement Durable et Biodiversité*.

Cette problématique transversale est donc interdisciplinaire et peut être abordée entre autres d'un point de vue biologique, mathématique et informatique. Chacune de ces disciplines apporte ses propres habitudes, contraintes et exigences synonymes d'enrichissement mais aussi vecteurs de conflits. Ces disciplines sont actrices des *Sciences d'Information Géographique* qui est en partie un facteur de cohésion scientifique de mes travaux de recherche.

On retrouve au sein de l'*UAG* une grande hétérogénéité dans l'utilisation des *SIG* qui sont présents dans de nombreux laboratoires (*DYNECAR*, *LAMIA*, *LARGE*, etc.) et dans différentes formations *Licence* et *Master Biologie Géologie Santé* (*BGS*), et *Master d'Informatique*.

Il apparaît nécessaire d'harmoniser l'utilisation de ces outils, notamment au travers d'enseignements transversaux, de développer les bonnes pratiques et de faire se rencontrer des communautés qui possèdent sans le savoir des problématiques connexes, des outils et approches complémentaires et des données insoupçonnées.

L'émergence de projets porteurs a en effet été le fruit de telles rencontres comme entre l'Informatique et la Biologie (projet *CESAR*) ou entre l'Informatique et la Santé (thèse de *Laurent Girdary*), mais cela pourrait aussi être le cas entre la Biologie et la Santé autour par exemple de l'étude des épidémies de dengue pour lesquels les facteurs environnementaux sont importants.

Cette volonté de partage d'expériences, de données, d'approches et de résultats peut s'appuyer *(i)* sur l'intérêt régional pour les *SIG* qui a réellement commencé en 2007 en Guadeloupe avec la création d'une *cellule SIG* au sein de la Région Guadeloupe, l'organisation de séminaires et de formations *(ii)* sur un rapprochement avec la Région Guyane qui possède de par son implantation et la station de réception *SEAS* une source potentielle de données immense *(iii)* sur la Région Martinique qui a mis en place le *SIG SIGMA* depuis 1992 et qui possède une expérience et des données non négligeables *(iv)* sur les partenaires locaux fournisseurs ou utilisateurs de données (*ONF, Parc National, DAF, BRGM, IRD*).

Fort de ces appuis, l'objectif dépasse les frontières des *Départements Français d'Amérique* (*DFA*) et vise le développement d'un réseau de *SIG* sur la Caraïbe afin de faciliter notamment les échanges de données en standardisant la définition des couches par exemple. L'une des applications visées est la prévention des risques autour des cyclones et des inondations dans la continuité du projet *PREVIOS* porté par différents laboratoires de l'*UAG* et ayant donné lieu à une thèse soutenue en 2010 (*Mathias Peroumalnaïk* encadré par *Gilles Enée*) ainsi que la thèse d'*Evelyne Fonseca-Cruz* en cotutelle avec l'Université de la Havane que je co-encadre depuis 2010.

L'idée principale est de pouvoir collecter facilement et rapidement une série de mesures faites avant, pendant et après le passage d'un cyclone sur les différentes iles du réseau afin d'affiner les modèles de prévision d'inondation par exemple. La principale difficulté des modèles prévisionnels sur l'impact des cyclones est le manque de données historiques du fait du faible (d'un point de vue apprentissage des modèles) nombre de passages d'un cyclone sur une île. Mais en tissant un tel réseau d'échange ce sont statistiquement une dizaine de phénomènes cycloniques par an qui pourraient alimenter une base commune d'apprentissage.

Par ailleurs, les problématiques de structuration et d'échange de données abordées dans ce cadre rejoignent mes perspectives dans le domaine de la modélisation des phénomènes diffus dont les cyclones sont un exemple.

Un dossier est en cours de montage dans le cadre du programme européen *INTERREG IV* et devrait être déposé courant 2012 avec des partenaires telles que *Cuba, Haïti, la Martinique, la Guyane,* ou *Saint-Domingue.*

6.2 Perspectives à court et moyen termes

Nous présentons dans cette section quelques perspectives relatives à des points techniques des contributions présentées dans la section 5. Ces perspectives concernent la troisième partie de la chaîne de traitement (Figure 1 page 29 et Figure 41 page 81) et les travaux relatifs à ces différents points sont en cours.

6.2.1 La sélection d'information

Les perspectives liées à la sélection d'information sont liées à la thèse de *Wilfried Segretier*. Les premiers résultats, issus du stage de *Frank Duhamel* en 2008, de recherches personnelles puis de la thèse en cours ont été le développement d'un algorithme de recherche permettant d'agréger les concepts d'une ontologie en respectant les relations sémantiques existant entre ces concepts.

Afin d'étendre les résultats obtenus, deux pistes sont envisagées pour les mois à venir. La première piste concerne l'algorithme mis en place. Certaines solutions renvoyées par l'algorithme sont très hétérogènes et composées de couches peu agrégées et d'autres réduites à un seul concept. Bien que cette dérive permette d'augmenter grandement la surface moyenne des unités, cela a pour effet de revenir sur le principe de fusion de couches qui était l'objectif visé au départ. Une solution conservant un minimum d'informations en provenance de chacune des couches est donc préférable. Pour cela nous définissons une compacité (δ) qui exprime l'écart maximal entre les niveaux de généralisation des concepts présents dans la couche agrégée. La Figure 42 illustre ce principe. Pour $\delta = 2$, si le concept C_3 (en bleu) est présent dans la couche d'information, les concepts C_8 et $C_4 - C_5 - C_6 - C_7$ (en vert) pourront l'être également mais le concept $C_8 - C_9$ (en rouge) est trop général ($d(C_3, C_8 - C_9) = 3$).

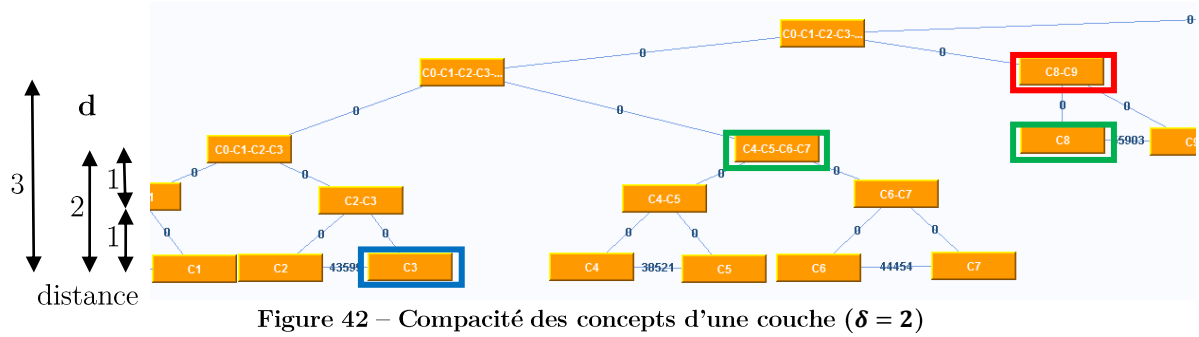
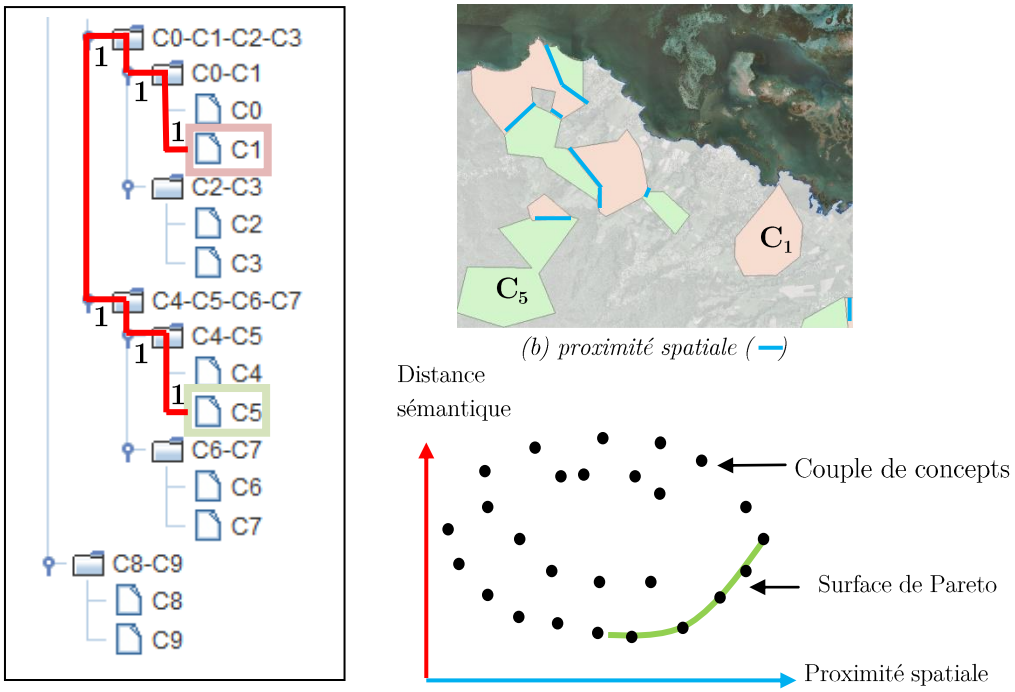


Figure 42 – Compacité des concepts d'une couche ($\delta = 2$)

La deuxième piste envisagée concerne l'utilisation d'une méthode de recherche par voisinages faisant évoluer une ou plusieurs solutions. La solution initiale est composée de tous les concepts feuilles des ontologies. Le voisinage est ensuite construit à partir de toutes les solutions accessibles en généralisant (agrégeant) deux concepts. Une solution du voisinage est ensuite évaluée en projetant le couple de concepts fusionnés dans l'espace (Proximité spatiale, Distance Sémantique) (pour une couche ayant n concepts, il y a $\frac{n(n-1)}{2}$ couples potentiels).

La distance sémantique est la distance entre les deux concepts dans l'ontologie. Son expression n'a pas encore été fixée mais elle doit faire intervenir la distance dans l'arbre ainsi que la différence de niveau de généralisation. La proximité spatiale est quant à elle proportionnelle à la longueur des frontières communes entre les instances de deux concepts. Ces deux distances sont a priori sans corrélations. La Figure 43, présente les deux critères utilisés (distance sémantique *(a)*, et proximité spatiale *(b)* ainsi que la projection des couples de concepts dans cet espace *(c)*).

La sélection dans le voisinage pourra alors se faire de plusieurs manières (à comparer) en sélectionnant par exemple aléatoirement une (ou plusieurs) solution parmi les solutions Pareto optimales vis-à-vis des deux critères pour choisir le successeur de la solution courante.



(a) distance sémantique (—) *(b) proximité spatiale (—)* *(c) projection des couples*

Figure 43 – Critères de l'approche par voisinage pour la sélection d'information

Ces deux critères génériques peuvent avoir plusieurs déclinaisons permettant d'intégrer par exemple le nombre d'instances d'un concept dans le calcul de la proximité spatiale.

Après validation et comparaison de cette approche avec l'approche par Algorithmes Génétiques, des tests sur une méthode d'agrégation localisée seront réalisés. En effet, les agrégations de concepts sont actuellement faites à l'échelle d'une couche et si deux concepts C_1 et C_2 sont généralisés en un concept C_3 toutes les instances de C_1 et C_2 sont rattachées à C_3. Ici nous proposons de faire une agrégation locale en fonction de paramètres spatiaux calculés sur les instances (adjacence, densité spatiale, etc.). On effectuera donc des agrégations par groupes d'instances et non par concept ce qui aura pour conséquence d'avoir dans une même couche plusieurs niveaux de généralisation.

L'apport d'une telle approche est l'aspect adaptatif du niveau de détail de l'information. La difficulté réside dans le choix des descripteurs géographiques et topologiques permettant de fixer le niveau d'abstraction local.

6.2.2 Affiner la classification vectorielle des forêts

Les paramètres environnementaux participant à la formation des écosystèmes, et plus particulièrement des écosystèmes forestiers, sont nombreux et interviennent dans différentes proportions. Notre approche vectorielle, permettant la classification des couverts forestiers de la Guadeloupe, se base sur des paramètres environnementaux importants mais qui peuvent être complétés et pondérés.

Nous pouvons nous appuyer pour cela sur plusieurs études qui intègrent différents paramètres pour aider la classification comme par exemple dans [**72**] où les auteurs utilisent des

paramètres climatologiques (outils *BIOCLIM*) ou dans [**114**] où les auteurs utilisent d'autres types de descripteurs (*Sobel*, etc.).

Par ailleurs, l'utilisation de modèles sur les classes cherchées permettrait de mieux gérer l'influence sur les écosystèmes des différents paramètres environnementaux utilisés. Dans [**55**] par exemple les auteurs utilisent un modèle statistique différent pour chaque classe de végétation et utilisent conjointement la probabilité associée à chacune des classes ainsi que son erreur standard de manière à éviter de classer une entité dans une classe fort probable mais peu fiable. Dans [**161**] les auteurs proposent un modèle de prédiction de la végétation en se basant sur la relation entre la distribution spatiale de la végétation de référence et des variables environnementales (modèle *PVM : Predictive Vegetation Modeling*). Le principe, déjà utilisé dans notre approche est de détecter les corrélations entre les échantillons d'observation et leur localisation. En effet, les deux informations ne sont pas indépendantes (continuité spatiale des paramètres environnementaux), ce qui vient appuyer le phénomène transitoire des formations forestières.

Ce dernier point soulève la question, encore ouverte, des dépendances spatiales dans les données biogéographiques. Nous avons indirectement introduit cette dépendance au travers de l'apprentissage sur les placettes de référence puisque nous avons intégré la latitude dans les données environnementales classantes. Mais la latitude est une information spatiale absolue qui peut être remplacée ou complétée par une information spatiale relative des Unités Vectorielles entre elles. L'intégration des relations spatiales dans le processus de classification étant un thème de recherche à lui seul, il sera détaillé dans la section 6.3.4.

6.2.3 Vers une classification 100% objet

L'ensemble des travaux menés jusqu'à présent en classification (calcul et sélection d'attributs, filtrage spatial, définition d'un dictionnaire, utilisation d'ontologies, approche vectorielle, etc.) nous conduit naturellement vers la mise en œuvre d'une approche 100% objet.

Historiquement la classification des images était d'abord basée sur une approche pixel [**211**], c'est-à-dire uniquement basée sur de l'information spectrale, sans prise en compte d'une information contextuelle (relations spatiales, classification des voisins, etc.). Mais la classification par une approche pixel, a montré des limitations ([**92**], [**168**]) notamment suite à l'apparition d'images à très haute résolution spatiale faisant apparaître des structures au sein même des objets (texture, etc.). Des zones homogènes d'un point de vu spectral représentant un même objet sont maintenant non homogènes car trop détaillées (information trop riche contenue dans les images) ce qui engendre un effet « poivre et sel » ([**196**], [**214**]) avec un certain nombre de pixels classés de manière isolée et un passage du raster vers le vectoriel très délicat. Cet effet, a effectivement été constaté lors de nos différentes expériences en classification pixel.

Comme indiqué dans [**214**] un certain nombre de méthodes ont été proposées pour atténuer cet effet comme des prétraitements (filtrage, analyse de texture), des classifications contextuelles ou des post-traitements (filtrage morphologique, etc.). L'inconvénient de ces différents filtrages est la perte de précision (flou artificiel aux frontières lié à l'utilisation de fenêtres d'analyse parfois très larges). Cette perte de précision engendre beaucoup de problèmes lors de la mise en correspondance des résultats de classification avec d'autres données vectorielles et doit être évitée.

Dans l'approche orientée objet, les éléments classés ne sont pas des pixels mais des objets comportant plusieurs pixels et ayant une certaine homogénéité spectrale. Mais on intègre également une information sur la géométrie de l'objet ou sur sa localisation [**127**]. La classification se fait alors en utilisant non seulement les attributs propres aux objets mais également en intégrant leurs relations spatiales ainsi que des contraintes d'intégrité entre les classes [**134**].

Cette approche prend tout son sens dans le contexte dans lequel nous nous plaçons puisque l'approche raster est d'ores et déjà couplée avec l'approche vectorielle et que des outils de caractérisation des textures et de définition vectorielle d'unités sont à disposition.

L'effort restant à faire est l'intégration des relations spatiales (section 6.3.4.) dans le processus de classification et la combinaison des vecteurs d'attributs issus de l'analyse des images avec les attributs vectoriels issus de la fusion des couches d'information. Deux approches, à comparer, sont possibles : *(i)* calculer les attributs image sur les *Unités Vectorielles* (UV) obtenues par fusion des couches *(ii)* obtenir des objets homogènes d'un point de vue spectral (par un algorithme de sur-segmentation tel que l'algorithme des bassins-versants [**132**]) et associer ensuite ces objets aux UV.

6.3 Perspectives à long terme

Les perspectives présentées ici sont celles associées à l'axe principal de mes recherches. L'idée majeure est d'étendre la modélisation des données et de l'intégrer dans la phase d'extraction de l'information. Nous abordons successivement : *(i)* une réflexion sur la modélisation des objets stricts imprécis et des objets diffus, *(ii)* l'intégration des erreurs de mesure dans les modèles en détaillant l'origine des erreurs, leur calcul et leur prise en compte *(iii)* une extension du modèle vectoriel flou présenté dans les contributions *(iv)* une réflexion sur l'intégration de la sémantique dans les données géo-spatiales *(v)* le problème du formalisme de la sémantique et enfin *(vi)* les relations spatiales et sémantiques.

6.3.1 Strict mais imprécis, diffus mais précis : vers un même modèle ?

Nous manipulons deux types d'objets au sein des *SIG* : des objets dont les frontières sont conceptuellement clairement définies mais dont la position peut être entachée d'erreurs et des objets dont les frontières ne sont pas précisément localisées (c'est le cas des formations forestières).

Dans l'approche raster on fait souvent appel à la théorie des ensembles flous pour représenter ou traiter les deux types d'objets [**112**] afin de traduire l'incertitude. La modélisation sous forme vectorielle de ces données ne doit à mon sens pas être faite de la même manière dans les deux cas et doit respecter le modèle conceptuel de l'objet puis nous devons proposer des outils ou paramètres adaptés permettant de traduire l'incertitude quant à la localisation de certaines des parties des objets. Cette position est appuyée par le fait que les erreurs de positionnement et les zones de transition entre phénomènes n'ont pas la même échelle.

Nous définirons donc d'une part un objet aux frontières bien définies mais peu précises par un objet strict annoté d'une information sur la qualité des positionnements. La section 6.3.1.1 détaille le mode de prise en compte envisagé pour les erreurs en expliquant leur origine et leur

calcul. Et d'autre part nous définirons dans la section 6.3.1.2., un phénomène transitoire en utilisant un modèle vectoriel flou qui est une extension de celui défini dans la section 5.3.3.

6.3.1.1 Des erreurs de mesure aux tolérances de positionnement

Comme le soulignent les auteurs de [**169**], les imprécisions liées aux mesures, à la localisation et à la caractérisation des données ne sont pas prises en compte dans la plupart des études bien qu'un certain nombre de paramètres exprimant la qualité des données géo-référencées puissent être renseignés [**203**] chacun déclinés dans la dimension appropriée (espace, temps, thématique).

Hors la prise en compte des erreurs dès leur apparition et leur quantification est importante car elles vont se propager lors des différentes opérations ensemblistes et ainsi générer un bruit multiplicatif sur l'information. Dans [**110**] les auteurs établissent un benchmark pour mesurer l'impact de ces erreurs de localisation sur des analyses telles que l'auto régression spatiale et montrent qu'elles sont loin d'êtres négligeables. Par ailleurs, la prise en compte des erreurs de précision dans les traitements est un sujet de recherche à part entière [**76**].

Les erreurs que nous voulons prendre en compte sont issues des méthodes d'extraction d'information à partir des données brutes (première partie de la chaîne de traitement). Elles dépendent des données utilisées, des méthodes d'extraction des attributs, et des performances des algorithmes de classification utilisés.

Dans notre cas, les données sources étant des images satellites ou aéroportées, une erreur de localisation liée à la résolution des images pourra être directement déduite des données. De la même manière, si les méthodes d'extraction d'attributs font appel à des fenêtres d'analyse, la taille de ces fenêtres induira une erreur de localisation (notamment aux frontières des objets).

D'un autre côté, les performances des outils de classification engendreront eux une erreur sur la sémantique (thématique) des objets, puisqu'une erreur de labellisation conduira à affecter un objet au mauvais concept dans l'ontologie.

Dans certains cas, par exemple lors de la partition de l'espace en différentes classes, traiter l'imprécision thématique et l'imprécision spatiale sont deux approches duales puisque la première est liée aux erreurs de localisation des frontières des classes donc à l'imprécision spatiale.

Il existe différents référentiels sur la précision des données géo-référencées et dont le principal, le *NSSDA* (*National Standard for Spatial Data Accuracy*) est essentiellement utilisé aux Etats-Unis. Ces référentiels considèrent que les erreurs de localisation suivent une distribution normale. Hors en 2008 Zandbergen [**217**] a caractérisé l'erreur de localisation pour un certain nombre de types de relevés (localisation GPS, géo-référencement des routes, etc.) ainsi que pour certaines bases de données standards (*TIGER roads*, *LIDAR*, etc.) et a fait apparaître des distributions plus complexes (distribution de *Rayleigh*, distribution logarithmique) non stationnaires.

Une étude sur la caractérisation des erreurs issues de la première étape de notre chaîne de traitement est donc complexe mais nécessaire afin de connaître plus précisément leur distribution.

L'ajout d'une information qualitative et quantitative sur la précision des mesures ou des traitements intervenus dans le processus de production d'une couche d'information permettrait de

résoudre un certain nombre de conflits lors des opérations ensemblistes en autorisant des modifications de la position des sommets des objets (ou des changements de leur sémantique).

L'objectif est donc d'associer aux concepts d'une couche d'information un degré de précision sur le positionnement et un taux de confiance sur leur sémantique. Le degré de précision donnera lieu à une tolérance sur le positionnement des sommets lors d'opérations ensemblistes afin de limiter la propagation de l'erreur. L'idée est de simplifier l'information (en éliminant des intersections) dès lors que la précision de sommets ne permet pas de justifier l'apparition d'un nouvel objet après une opération telle que l'intersection. Les règles de propagation du degré de précision et du taux de confiance devront être définies à partir de développements théoriques. La Figure 44 illustre de principe de correction, partant de deux objets provenant de deux couches différentes ayant une tolérance définie pour chaque sommet (symbolisé par les disques entourant les sommets) (a), on réalise l'intersection des deux couches. Sans rectification *(b)*, on constate une zone d'intersection entre les deux objets. En utilisant la tolérance *(c)* la zone d'intersection disparaît dans la mesure où les sommets d'un objet sont inclus dans la zone de tolérance de l'autre objet. La figure illustre également que l'erreur quant à la position des sommets modifiés change (cercles de taille différente autour des sommets modifiés) sans pour autant que l'illustration soit représentative de la nouvelle erreur qui est encore une fois à quantifier de manière théorique.

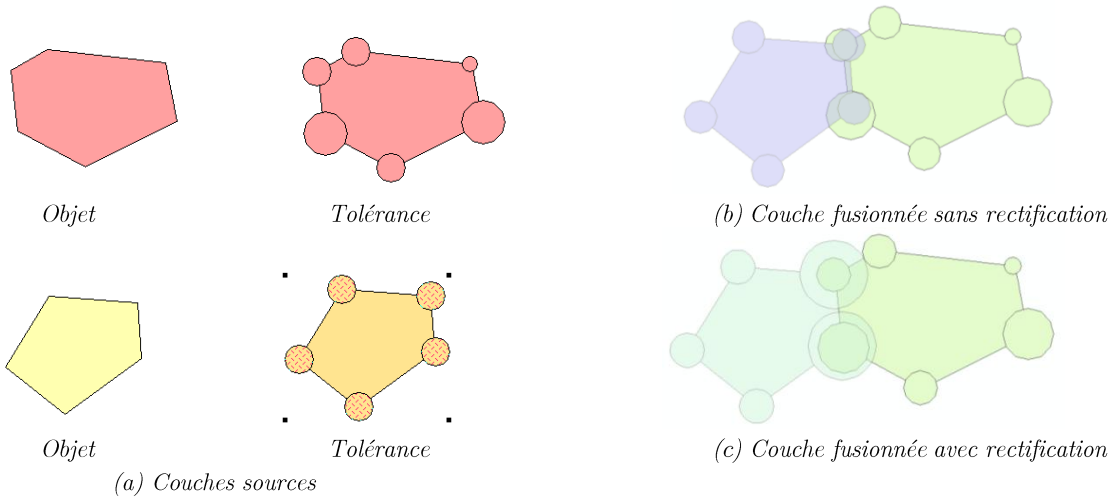


Figure 44 – Tolérance de positionnement

6.3.1.2 Extension du modèle vectoriel flou

Le passage à une véritable représentation vectorielle floue nécessite de redéfinir les objets élémentaires (point, ligne, surface) et les opérations ensemblistes (union, intersection, etc.) présents dans les SIG. Ce passage ne peut se faire qu'en plusieurs étapes en s'appuyant par exemple sur des modèles comme celui présenté dans [**67**] où les auteurs représentent un objet flou par trois objets stricts : l'intérieur, la frontière et l'extérieur (L'extérieur représentant la zone d'incertitude). C'est une vue restrictive des objets qui a néanmoins le mérite de simplifier les traitements et analyses. Notre approche a d'ores et déjà poussé un peu plus loin cette représentation en ne distinguant pas ces trois parties et en proposant différents zones intermédiaires. Mais certains aspects des ensembles flous sont encore manquants dans notre modèle comme la notion d'ensemble ouvert ou de gradient local de la fonction d'appartenance.

La Figure 45 montre la représentation d'un ensemble vectoriel flou tel que nous l'envisageons dans un premier temps. La différence avec celle obtenue dans la section 5.3.3 est l'ajout d'un vecteur, à chaque sommet des différentes zones tampons, indiquant le gradient local de la fonction d'appartenance. La combinaison des deux champs de vecteurs représentés sur la Figure 45 permettra de renseigner un peu plus la zone tampon qui était jusque là vue comme une zone homogène. Lors d'opérations ensemblistes sur de tels modèles, on peut imaginer déformer les contours d'une zone tampon en tenant compte de ce gradient pour mettre en correspondance des frontières. Pour cela, on peut s'inspirer des travaux sur les contours déformables (snake [**87**]) et sur les travaux de [**78**] intégrant des relations spatiales dans les modèles déformables. La redéfinition des opérations ensemblistes dans un tel contexte nécessite de redéfinir formellement tous les objets et toutes les opérations en mettant en place une algèbre particulière. Des outils de visualisation permettant de représenter de manière multi-échelle ces champs de gradient doivent également être proposés. Tout ceci doit être fait avec précaution pour assurer notamment la continuité de la fonction d'appartenance après les différentes opérations ensemblistes et garantir en même temps une meilleure fiabilité dans la représentation des données.

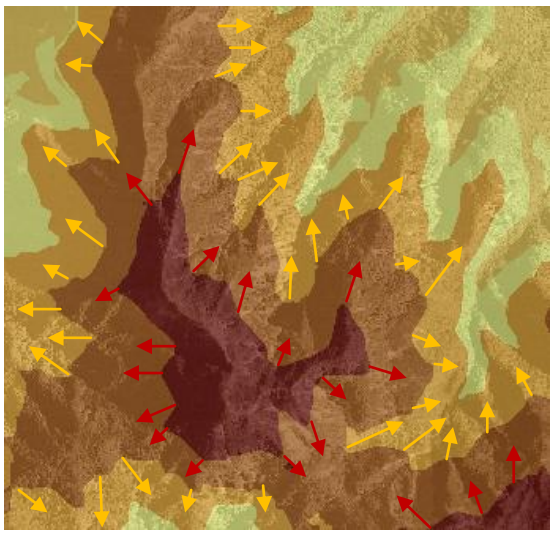


Figure 45 – Représentation du gradient de la fonction d'appartenance

Par ailleurs, dans le modèle actuel que nous avons proposé, les zones tampon définissant une vue v d'un ensemble diffus sont délimitées par des frontières permettant de regrouper les valeurs de la fonction d'appartenance en un certain nombre d'intervalles ($I^v = \{ I_0^v, I_2^v, \dots, I_{g_v}^v \}$, section 5.3.3 page 65). Les zones tampon et les frontières ne contiennent que les valeurs des bornes des intervalles. Hors chaque sommet est à l'origine positionné sur une *Unité Vectorielle* (*UV*) qui possède une valeur pour la fonction d'appartenance. Ceci donne la possibilité de transmettre la valeur de la fonction d'appartenance de cette *UV* sur le sommet correspondant. Nous introduisons ainsi une information supplémentaire conduisant à un gradient de la fonction d'appartenance le long de la frontière.

Les structures de données actuelles ne permettent pas d'intégrer facilement des données tabulaires à l'échelle des sommets d'un objet complexe. Une étude est donc à mener pour trouver un moyen d'intégrer ces informations (valeur de la fonction d'appartenance et vecteur gradient) sans alourdir la structure de donnée.

Pour pousser encore plus loin la modélisation des phénomènes diffus nous envisageons de les représenter en partie à l'aide de polygones ouverts. Cette perspective est inspirée des travaux de *Pascal Matsakis* de l'*Université de Guelph* au *Canada* avec qui des échanges ont commencé en 2011. Il s'est intéressé à la représentation de l'information directionnelle sous forme raster [**160**] puis vectorielle [**83**]. L'objectif est ici d'adapter le principe de calcul des polygones ouverts afin de représenter le gradient des fonctions d'appartenance.

Ici le plan est séparé en régions (imbriquées ou disjointes) représentées par des polygones ouverts qui permette de partitionner le plan en intervalles de valeurs pour la fonction d'appartenance. La Figure 46 donne une illustration d'un polygone ouvert issue de [**83**]. Cette perspective nécessite la définition d'un type abstrait de donnée (demi-lignes) pour la représentation de polygones ouverts.

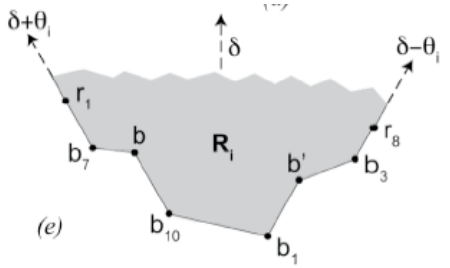


Figure 46 – Polygone ouvert (extrait de [83])

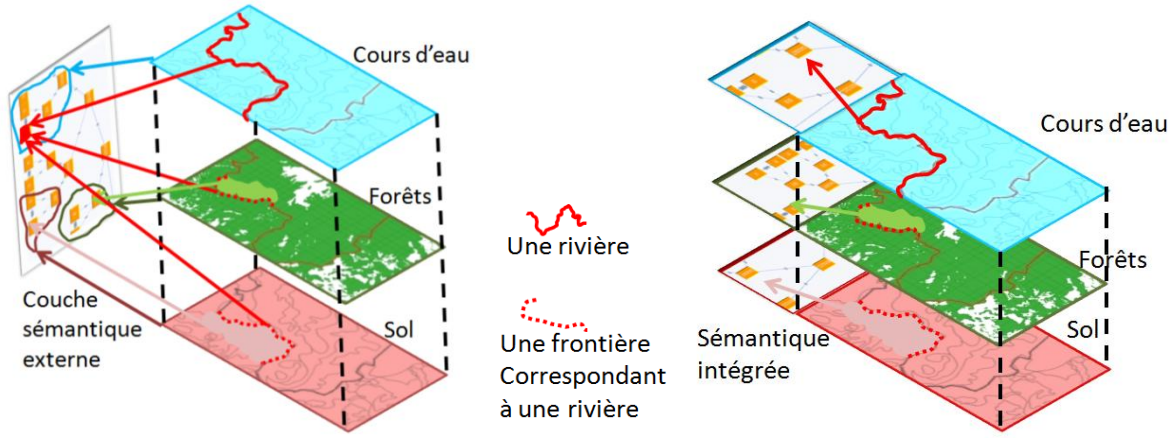
6.3.2 Couche sémantique externe ou sémantique intégrée ?

Cette question semble être, de mon point de vue, la plus ouverte parmi mes perspectives. Les deux approches ont été partiellement abordées puisqu'un modèle de couche sémantique partagée a été proposé (section 5.3.4 page 72) et que l'intégration d'une sémantique au sein d'une couche a été mise en œuvre en incluant des références à des concepts d'une ontologie dans le processus de sélection d'information (section 5.4.1 page 75).

Dans la définition d'une couche sémantique externe, deux verrous se présentent. Le premier concerne la définition même du modèle et des contraintes d'intégrité entre les concepts qui peuvent s'appuyer sur les relations spatiales [**195**] (section 6.3.4) et sémantiques [**152**] (section 6.3.4.2) existantes entre les concepts. Le deuxième est posé par la non ou mauvaise utilisation des modèles relationnels entre tables dans les *SIG*. En effet, bien qu'étant des *Systèmes d'Information* (*SI*) les *SIG* ne sont pas axés sur les schémas relationnels entre tables et les tables relationnelles ne sont généralement utilisées que ponctuellement lors de jointures sémantiques ou spatiales par exemple mais pas dans la définition même des données. En effet, d'après [**61**] les données géo-référencées sont souvent mal structurées et ne respectent pas les règles en vigueur dans les *SI*. On ne trouve pas de processus de construction d'une couche et l'information sémantique permettant de mettre en relation différentes couches est souvent présente mais non identifiée. On trouve néanmoins des travaux relatifs à cette approche dans la littérature puisque dans [**204**] et [**60**] les auteurs proposent une couche sémantique ontologique intermédiaire entre l'utilisateur et les données pour interroger une base de données spatiale. Dans [**61**], les mêmes auteurs utilisent une localisation indirecte des objets de manière à ce que chaque table fasse référence à la même topologie. Ceci alourdit grandement les traitements du fait des nombreuses

références vers d'autres tables mais est conceptuellement beaucoup plus rigoureux et plus proche de ce que l'on trouve dans les *SGBD* classiques (clé externe, table d'association, etc.).

La difficulté de cette approche réside donc en partie dans une démarche de conception des couches et dans la définition des relations spatiales (section 6.3.4) et sémantiques (section 6.3.4.2) entre concepts dans la couche sémantique ainsi que la définition d'un schéma relationnel standardisé entre les couches et la couche sémantique. L'avantage est une mise en correspondance constante des objets permettant de lever toute ambiguïté lors d'un croisement de couches. Dans la Figure 47 (a) les trois couches thématiques sont associées à trois sous-arbres d'une même ontologie. Les objets présents dans ces couches sont annotés par des concepts respectifs de ces sous-arbres. Et on peut également annoter une partie d'un objet à partir par un concept d'un sous-arbre associé à une autre couche. C'est le cas pour les couches *sol* et *forêt* qui ont une frontière définie par une rivière (en pointillé rouge) qui est annotée par un concept rattaché à la couche *cours d'eau* (flèches rouges).



(a) Couche sémantique externe *(b) Couche augmentée*

Figure 47 – (a) Couche sémantique externe - (b) Couche augmentée

Concernant l'intégration de la sémantique au sein même d'une couche, l'ontologie sera stockée en utilisant les métadonnées descriptives qui sont instaurées dans les *SIG* et dont il faudra formaliser le contenu (section 6.3.3).

L'inconvénient de cette approche est qu'il n'y a pas la possibilité d'annoter un objet avec un concept d'une autre ontologie que celle de la couche. Dans la Figure 47 (b) il n'y a pas de possibilité d'annoter la frontière représentant une rivière avec le concept *rivière* de la couche sur les cours d'eau. Cette approche nécessite donc des mécanismes de mise en correspondance d'ontologies lors d'un croisement de couches.

Les deux approches nécessitent d'ajouter des champs dans les tables pour que chaque objet fasse référence à un concept de l'ontologie et que les attributs liés au concept soient renseignés.

6.3.3 De la nécessité d'une standardisation à l'utilisation d'un formalisme : des outils orientés utilisateur

La tolérance sur la position des sommets et l'utilisation de la sémantique ont le même objectif, permettre la mise en correspondance de plusieurs représentations d'une même entité en limitant les erreurs liées au positionnement. Les deux approches peuvent être combinées par exemple en définissant lorsque cela est possible des points d'ancrage sur les données. Ces points d'ancrage sont rattachés à un concept (ou encore mieux une instance) et permettent de résoudre certains conflits. Les autres parties des objets, non annotées, pourront alors bouger autour de leur position en fonction de la tolérance qui leur est attribuée pour être mises en correspondance avec d'autres parties d'objets (points d'ancrage ou non).

Mais l'ensemble de ces perspectives ne seront possibles qu'en standardisant la représentation des ontologies et des erreurs au sein des *SIG* au travers notamment des méta-données descriptives. Les méta-données concernant les informations spatiales sont normalisées par les normes *ISO 19115* et *ISO 19000* mais celles-ci ne concernent que les attributs géographiques (référentiel, système de coordonnées, etc.) et non les informations contextuelles (source des données, structuration de l'information, information qualitative, etc.). Dans [**187**] les auteurs définissent 8 catégories pour permettre de décrire l'ontologie sur les données : méthode de collecte, définition des termes, Système de mesure (seuils, bornes, utile pour les comparaisons de données), Système de classification, modèle de données, raisons de la collecte, contraintes, commentaires. Mais les méta-données inclues de cette manière ne sont pas formalisées.

Nous envisageons donc dans nos perspectives de formaliser une partie de ces informations de manière à permettre leur utilisation par des outils dédiés aux ontologies. Plusieurs langages sont disponibles pour nous aider dans cette tâche. L'*OGC* (*Open Geospatial Consortium*) fournit beaucoup de standards pour les *SIG* et de langages basés sur la technologie *XML* : *(i) GML* qui est un langage pour décrire les données géographiques *(ii) OWL* (*Ontologie Web Language*) pour formaliser les ontologies *(iii) ONTOAST* [**163**] qui est une union d'*AROM-ST* et d'*AROM-OTON*, deux extensions d'*AROM* pour la gestion des données spatio-temporelles et des ontologies à la manière d'*OWL*. Ces derniers sont plus adaptés aux données spatiales qu'*OWL* qui est un outil très généraliste *(iv) KML* pour la définition des opérateurs spatiaux [**76**] *(v) OCL* pour le codage des contraintes d'intégrité.

Mais ces outils, bien que disponibles ne sont pas intégrés dans les *SIG* actuels [**57**] dans le sens où il n'existe pas d'outils permettant d'interfacer les ontologies malgré le fait que dès 2002 Fonseca ([**101**], [**102**]) puis Cruz en 2005 [**85**] proposent un *SIG* guidé par les ontologies (*Ontology driven GIS*). L'association entre les données géo-référencées et les ontologies est faite dans des outils externes tels que *GeoSVM* ([**101**], [**102**]) et *SPIRIT* [**122**]. L'inconvénient est que cette association reste réservée à des spécialistes (80% des utilisateurs d'ontologies sont dans des institutions de recherche selon [**73**]) et que les principaux utilisateurs responsables de la collecte des données géo-référencées ne les utilisent pas.

De manière à rendre exploitables nos outils et dans la mesure où le grand public requiert des outils simples de compréhension et d'utilisation [**57**] nous nous limiterons dans un premier temps à formaliser les ontologies dans les méta-données et à définir des outils intégrables aux *SIG* permettant d'interagir avec les ontologies présentes dans les méta-données des différentes couches. Ces outils devront permettre la navigation et la mise en correspondance des concepts et des instances.

6.3.4 Les relations entre objets et concepts

6.3.4.1 Les relations spatiales

La définition des relations spatiales entre concepts ou instances est un vaste domaine de recherche auquel nous avons régulièrement fait référence au cours notamment des perspectives précédentes. Leur intégration est nécessaire à plusieurs niveaux dans nos approches : dans la couche sémantique pour décrire de manière abstraite une scène ; dans les algorithmes de sélection d'information afin de guider les recherches ; dans la classification orientée objet ; dans la modélisation des types de forêts et de leur relations. L'intégration des relations spatiales dans nos modèles est envisagée en s'inspirant des travaux de *Tarquini* [**195**] ou de *Clementini* [**76**] qui proposent un cadre conceptuel pour modéliser les relations spatiales en donnant différents niveaux de représentation ainsi que des attributs sur les relations comme la cardinalité ou la granularité.

Les relations spatiales ont été définies par *Eigenhofer* en 1991 ([**93**], [**94**]) avec notamment un formalisme décrivant tous les types de relations existants entre différents types d'objets (lignes, surfaces, etc.) faisant intervenir des relations telles que *disjoint*, *intersecte*, *touche*, etc. Ces relations spatiales ont été étendues notamment par *Bloch* et *Matsakis* ([**68**], [**153**], [**158**]) à des représentations floues ou incertaines permettant de décrire des relations telles que *proche de, au nord de* ou *touche peut être*. Ces travaux sont allé jusqu'à la définition d'ontologies décrivant les relations spatiales floues ([**119**], [**57**]).

En fonction du niveau d'abstraction voulu, différentes approches sont possibles pour exprimer les relations spatiales. Dans le cas de la sélection d'information ou de la classification orientée objet, on se place dans l'espace des instances (objets) qui disposent d'une localisation définie et pour lesquelles il est possible de quantifier les relations spatiales. Par exemple la relation *intersecte* peut être quantifiée par la surface d'intersection et la relation *touche* par la longueur de la frontière commune. D'autres relations plus abstraites, *au nord de* par exemple, peuvent également être quantifiées en utilisant une information directionnelle calculée à partir de *F-Templates* ou de *F-Histogrammes* (Histogramme de Force entre autre) définis dans [**160**], [**82**].

Dans ce cadre, un outil permettant de calculer le graphe d'adjacence a d'ores et déjà été développé (Figure 48), le graphe est ici valué en fonction de la longueur de l'arrête commune entre les deux objets reliés.

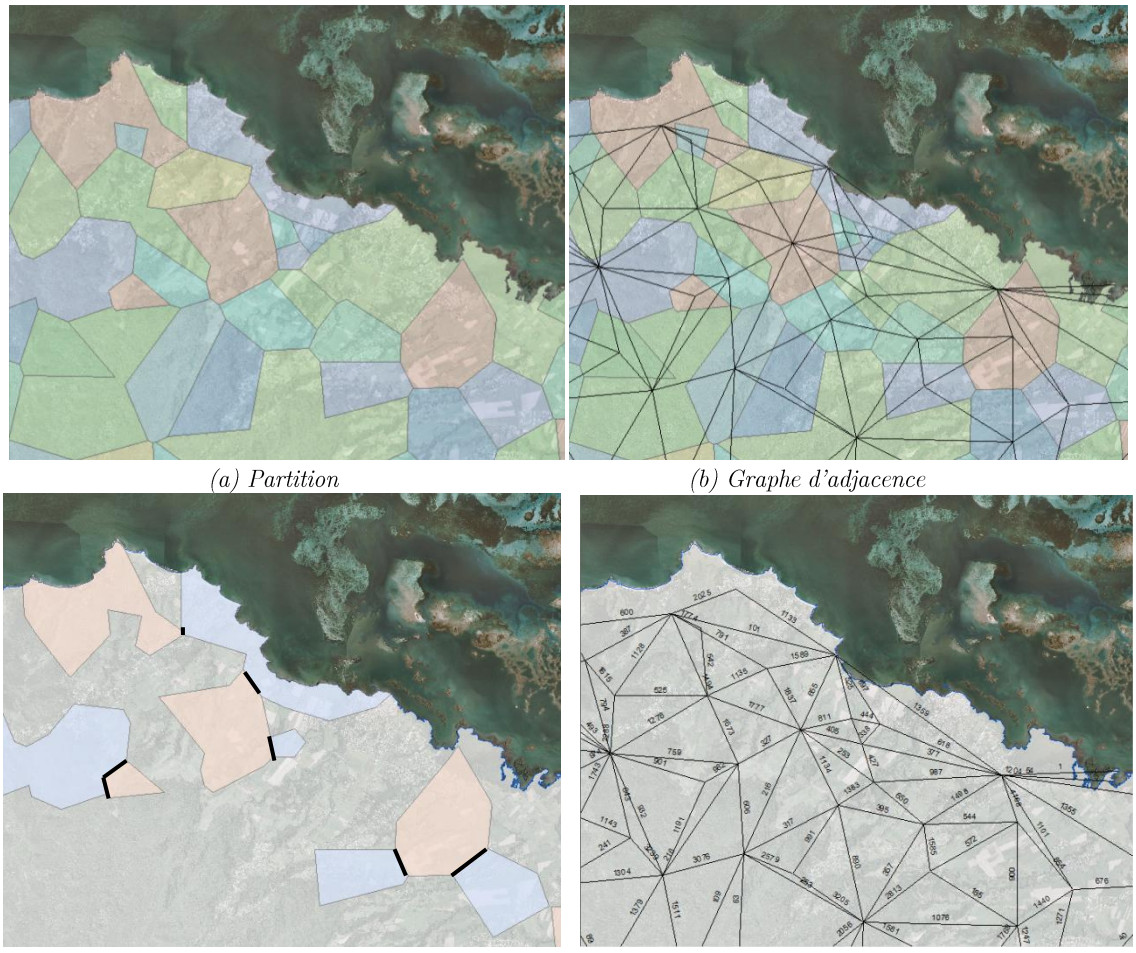
Les différentes étapes précédentes font appel à de la géométrie (*diagramme de Voronoï*), des calculs ensemblistes (*intersection d'ensembles, connexité*) et des statistiques.

Dans le cadre plus abstrait de la couche sémantique ou de la modélisation des forêts, les relations spatiales entre concepts doivent généraliser les relations entre les instances. Deux approches complémentaires sont envisagées.

1. La première est une approche descendante consistant à définir dans un premier temps les concepts et les relations spatiales (*proche de, disjoint, au nord de*) de manière formelle (approche conceptuelle). Cette définition permet également de définir des contraintes d'intégrité entre les concepts (disjoint, connexe, etc.). Dans un second temps, les relations spatiales peuvent être quantifiées pour une instance de la couche correspondant. La prise

en compte des contraintes d'intégrité peut également permettre de corriger certaines erreurs de positionnement des objets.

La deuxième approche est une approche ascendante permettant de définir les relations à partir d'une analyse directe des couches d'information : on construit le modèle à partir des instances. Nous nous appuyons notamment dans ce cadre sur les travaux de *Baglioni* [**61**] qui présentent une méthode d'extraction automatique d'ontologies et de relations spatiales à partir d'une base de données spatiale ainsi que sur ceux de *Guesgen* [**112**] pour la quantification des relations spatiales. Cette approche est plus directe et plus facile à mettre en œuvre mais le modèle construit dépend fortement de l'instance de la couche ce qui peut poser certains problèmes, notamment dans l'interprétation des relations.



(a) Partition *(b) Graphe d'adjacence*

(c) adjacence de deux classes C_1 (Rose) et C_2 (bleu) *(d) Graphe d'ajdacence valué*

Figure 48 – Construction du graph d'adjacence valué

6.3.4.2 Les relations sémantiques

Les relations sémantiques entre objets sont directement liées à la manière de structurer les concepts manipulés. Il faut donc commencer par définir des ontologies spécifiques pour les différentes couches manipulées avant de pouvoir définir les relations sémantiques entre les concepts (ou objets, un objet étant une instance d'un concept). On peut se baser pour cela sur les travaux de *Deliiska* [**89**] qui définit des ontologies dans le cadre des données spatiales ainsi que sur les travaux de *Raskin* [**178**] qui travaille plus spécifiquement sur la découverte de sémantique

dans les données spatiales environnementales. De même dans [**61**] et [**60**], *Baglioni* définit un moyen de construire une ontologie à partir d'une base de données spatiale. Cette étape nécessite une expertise et une connaissance de la taxonomie des domaines liés à chacune des couches. Des outils comme GSA (Geospatial Semantic Analytic Framework [**59**]) ou *SWEETO-GS* qui est une ontologie sur les données spatiales sont à prendre en compte pour définir les ontologies propres à nos données. Une étude a été menée dans ce cadre en 2010 pour définir une ontologie sur les couverts forestiers en collaboration avec le laboratoire *DYNECAR* dans le cadre d'un stage de Master 2ème année.

La déclinaison de l'étude des relations sémantiques entre concepts dans le cadre des géo-ontologies (ontologies spatiales) n'est pas simple. Dans notre cas, au sein d'une ontologie associée à une couche thématique, les relations sémantiques existantes seront dans un premier temps essentiellement de type *est-un,* permettant ainsi de structurer les ontologies sous forme d'arbre de concepts avec différents niveaux de généralisation. Les relations plus complexes entre concepts d'un même niveau seront donc déduites et quantifiées (notion de distance sémantique) en se basant sur la distance dans l'arbre (c'est le principe déjà utilisé dans la sélection d'information section 836.2.1). D'autres approches seront à prendre en compte pour définir des métriques plus fines.

Par ailleurs, une des problématiques soulevées par l'introduction d'une sémantique au sein des couches est celle de la mise en correspondance de d'ontologies provenant de couches thématiques différentes ([**128**]). Des méthodes et des outils pour rapprocher des ontologies ont été proposées par différents auteurs ([**103**], [**128**], [**180**]) mais ces approches très générales sont à décliner dans le cadre géo-spatial en limitant le spectre de certaines ontologies. Par exemple, en définissant une ontologie spécifique pour l'annotation des frontières des objets qui peuvent être partagées par plusieurs couches d'information.

Enfin, l'intégration des contraintes d'intégrité dans les relations sémantiques avec notamment les travaux de *Mäs* ([**150**], [**151**], [**152**]) rejoint la problématique de la définition des relations spatiales entre concepts. Ces deux problématiques devront donc être traitées conjointement.

6.4 Schéma synthétique

Ces nombreuses perspectives sont ambitieuses et il est évident qu'elles s'inscrivent dans le long terme et ne pourront être toutes traitées simultanément. Elles seront l'objet de constantes remises en question en fonction de l'évolution des *SIG* et des contributions des nombreux chercheurs travaillant dans les domaines cités. Par ailleurs, elles ne seront, et ne doivent, pas être un travail solitaire. Elles s'appuieront donc en partie sur des encadrements de stages de *Master* et de *doctorat*s dont deux thèses notamment en cours, *Eveline Fonseca-Cruz* et *Wilfried Segretier*, sur certaines de ces perspectives. Elles s'appuieront aussi sur les collaborations en cours ou à venir avec des partenaires locaux, nationaux et internationaux.

Le schéma de la Figure 49 synthétise l'objectif visé par l'axe de recherche dans lequel je me suis engagé depuis 2008. Il regroupe mes contributions et mes perspectives. La finalité est d'obtenir un passage des données brutes raster vers une information vectorielle exploitable dans un contexte intégrant sémantique, modélisation, relations spatiales, etc. de manière transparente.

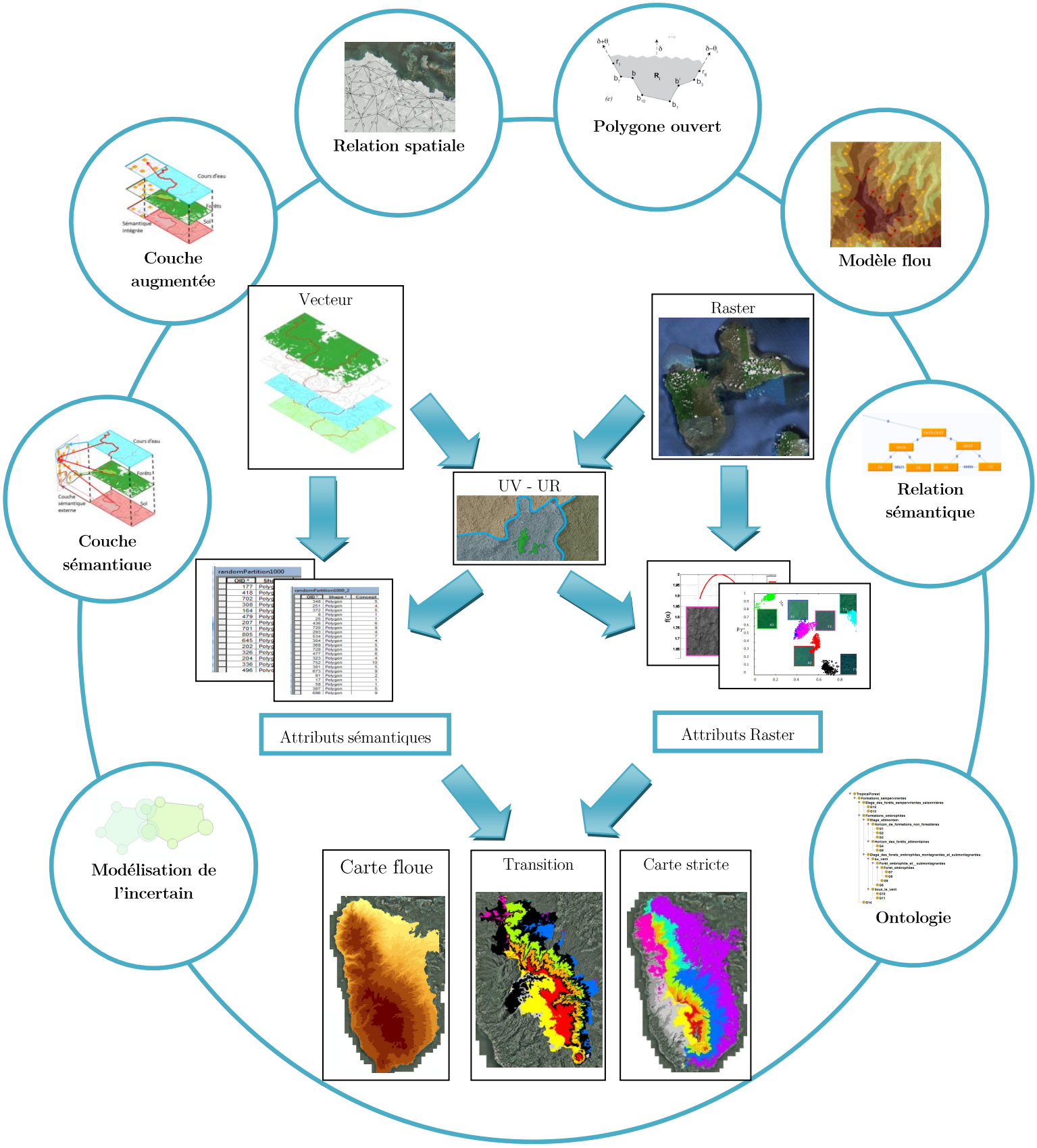


Figure 49 – Synthèse des contributions et perspectives

7 Conclusion: une projection au milieu des défis des *SIG*

Dans son article *Twenty Years of Progress : GIScience in 2010* [**108**] *Goodchild* fait une rétrospective des avancés des *SIG* et de leurs enjeux, notamment en recherche, dans les dix prochaines années. Cet article, bien que partiellement subjectif conforte l'axe de recherche sur lequel je me suis engagé depuis 2008. En effet, le foisonnement de disciplines, de méthodologies et de technologies gravitant autour des *SIG* est un catalyseur pour un certain nombre de recherches qui sont nées ou ont pris forme grâce à ce contexte. De manière très globale, l'interopérabilité est au cœur des préoccupations actuelles et il n'est pas étonnant d'en retrouver une déclinaison au sein des *SIG* qui sont historiquement basés sur des problématiques et outils transversaux.

La Figure 50 est le résultat d'une étude bibliographique faite par *Fisher* en 2006 [**96**], elle montre l'importance des auteurs (en termes de publication et de référencement) et leur proximité thématique. Des noms comme *Burrough*, *Goodchild* ou *Egenhofer* se démarquent naturellement du fait de leurs contributions dans la définition des modèles de données, sur lesquels les *SIG* sont basés, ou dans l'analyse spatiale qui a été l'un des grands chantiers des deux dernières décennies.

Même si les thématiques de recherche des auteurs se recoupent, la répartition spatiale dans la Figure 50 suit la logique suivante : en *haut à gauche* sont représentés les auteurs ayant travaillé sur la *modélisation de l'environnement*, en *haut à droite* sur la *modélisation et la représentation des données* ainsi que sur les *ontologies*, en *bas à gauche* sur la *télédétection* et *l'incertain,* en *bas au centre* sur *l'analyse spatiale* et en *bas à droite* sur des *règles de décision.*

Si je devais projeter mon parcours dans cette représentation il suivrait la courbe affichée en rouge dans la figure : historiquement parti de la télédétection et avec comme supports des applications en environnement c'est donc vers la modélisation des données que je m'oriente depuis 2008. L'originalité de mes contributions et perspectives dans ce schéma est entre autre la modélisation sémantique et vectorielle d'objets diffus. Ceci reviendrait à créer un sixième ensemble regroupant *incertain* et *modélisation de données* avec une déclinaison vectorielle dans lequel nous pourrions projeter, en plus de ceux déjà présents, des noms tels que *Bloch* ou *Matsakis.*

Enfin, les perspectives que nous avons détaillées dans ce mémoire sont également confortées par les défis à relever soulignés par *Goodchild.* En effet, ces défis rejoignent l'idée de la modélisation et de l'intégration des erreurs en considérant le caractère incertain des données et des traitements (jointure spatiale) ainsi que l'idée d'intégration de la sémantique dans les couches d'information. Il indique par ailleurs l'importance de la mise en correspondance des couches d'informations en indiquant par exemple qu'il est beaucoup plus facile d'accéder à une information sur l'ensemble d'une couche que d'accéder à un ensemble d'information (provenant de plusieurs couches) en un emplacement précis. Ceci est dû au modèle conceptuel des données dans un *SIG* qui structure l'information en couches thématiques. La modélisation sémantique d'une couche à laquelle les couches d'instance font référence est un moyen de résoudre partiellement ce problème.

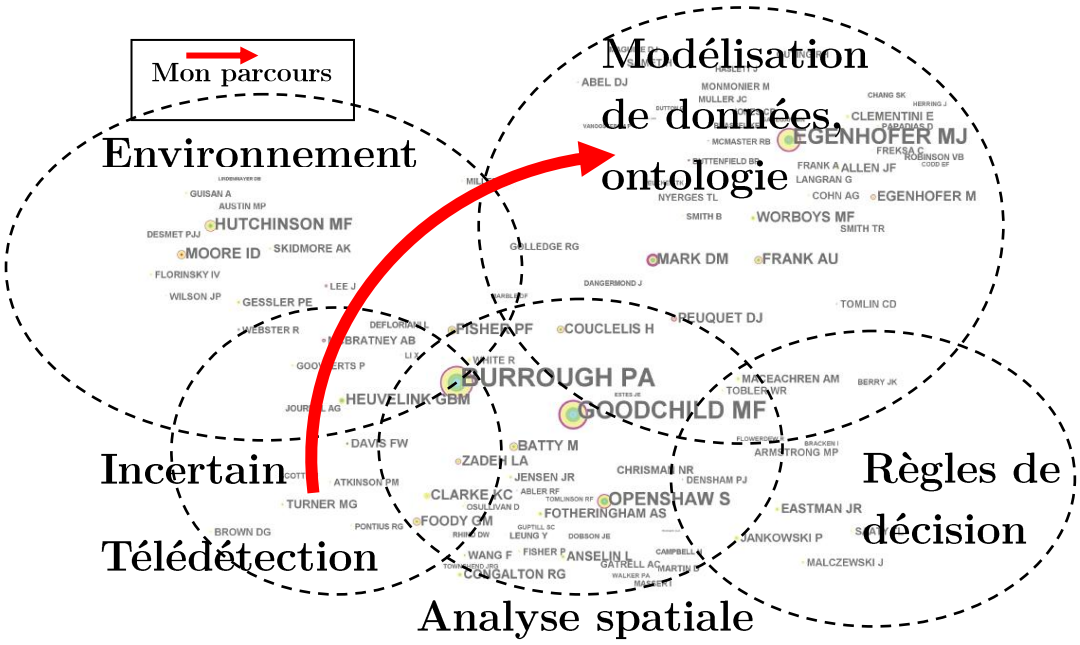


Figure 50 – Répartition thématique des auteurs basée sur un extrait de [108]

8 Annexes

8.1 Rapports

[43] **Grandchamp, E.,** *Dossier de projet et rapports d'exécution du projet CESAR.* Programme Européen INTERREG IIIb Espace Caraïbes. 2002, 2005, 2006, 2007

[44] **Grandchamp, E.,** *Dossier de projet et rapports d'exécution du projet CESAR Volet II.* Programme Européen INTERREG IV. 2010, 2011

[45] **Grandchamp, E.**, *Rapport d'exécution dans le cadre du projet PARAGE*, 2007

[46] **Grandchamp, E.,** *Modélisation et algorithme de résolution.* Rapport Interne Alcatel Space Industries. 2000

[47] **Grandchamp, E.,** *Classification des orbites.* Rapport Interne Alcatel Space Industries. 2000

[48] **Grandchamp, E.,** *Optimisation de constellations par approche géométrique.* Rapport Interne Alcatel Space Industries. 2000

[49] **Grandchamp, E.,** *Etude de la méthode de Hanson.* Rapport Interne Alcatel Space Industries. 1999

[50] **Grandchamp, E.,** *Etude de FlexToolGA Partie I, II, III et IV.* Rapport Interne Alcatel Space Industries. 1999

[51] **Grandchamp, E.,** *Etude des doublets de satellites.* Rapport Interne Alcatel Space Industries. 1999

[52] **Grandchamp, E.,** *Modélisation mathématique des critères.* Rapport Interne Alcatel Space Industries. 1999

[53] **Grandchamp, E.,** *Bibliographie sur les méthodes d'optimisation.* Rapport Interne Alcatel Space Industries. 1998

[54] **Grandchamp**, **E.,** *Optimisation de constellations de satellites : première approche.* Rapport Interne Alcatel Space Industries. 1998

9 Références

Les références de **[37]** à **[54]** correspondant aux publications personnelles (revues, conférences, rapports) sont présentées aux pages 21 et 99.

[55] Mohamed Abadi, "Couleur et texture pour la classification d'images satellites haute résolution," Université des Antilles et de la Guyane, Ph.D. dissertation 2008.

[56] Arnon Accada and David T. Neilb, "Modelling pre-clearing vegetation distribution using GIS-integrated statistical, ecological and data models: A case study from the wet tropics of Northeastern Australia," *Ecological Modelling*, vol. 198, pp. 85-100, 2006.

[57] A. Al-Ani, M. Deriche, and J. Chebil, "A new mutual information based measure for feature selection," *Intelligent Data Analysis*, vol. 7, no. 1, pp. 43-57, 2003.

[58] Jochen Albrecht, Brandon Derman, and Laxmi Ramasubramanian, "Geo-ontology Tools: The Missing Link," *Transactions in GIS*, vol. 12, no. 4, pp. 409–424, 2008.

[59] D. Altman, "Fuzzy set theoretic approaches for handling imprecision in spatial analysis," *International Journal Geographical Information Systems*, vol. 8, no. 3, pp. 271-289, 1994.

[60] Ismailcem Budak Arpinar et al., "Geospatial Ontology Development and Semantic Analytics," *Transactions in GIS*, vol. 10, no. 4, pp. 551-575, 2006.

[61] Miriam Baglioni and al., "Ontology-supported Querying of Geographical Databases," *Transactions in GIS*, vol. 12, pp. 31-44, 2008.

[62] M. Baglioni, M. V. Masserotti, C. Renso, and L. Spinsanti, "Building geospatial ontologies from geographical databases," *GeoSpatial Semantics: Proceedings of the Second International Conference (GeoS 2007). Berlin, Springer Lecture Notes in Computer Sciences*, vol. 4853, pp. 195–209, 2007.

[63] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behavioral Sciences*, vol. 12, no. 2, pp. 153-155, March 1967.

[64] Base de données UCI Repository : http://archive.ics.uci.edu/ml/datasets.html.

[65] Jane Bemigisha et al., "Representation of Uncertainty and Integration of PGIS-based Grazing Intensity Maps Using Evidential Belief Functions," *Transactions in GIS*, vol. 13, no. 3, pp. 273-293, 2009.

[66] Ursula C. Benz and al., "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS Journal of Photogrammetry \& Remote Sensing*, vol. 58, pp. 239-258, 2004.

[67] N. Bioret, Guillaume Moreau, and M. Servières, "Géolocalisation en milieu urbain par appariement entre une collection d'images et un SIG 2D," *Ingénierie des Systèmes d'Information, numéro spécial Systèmes d'information et géo-localisation*, vol. 14, no. 5, pp. 107-131, 2009.

[68] Jan T. Bjorke, "Topological relations between fuzzy regions: derivation of verbal terms," *Fuzzy Sets and Systems*, vol. 141, pp. 449–467, 2004.

[69] Isabelle Bloch, Olivier Colliot, and Roberto M. Cesar, "Modélisation de la relation spatiale « entre » pour des objets d'extensions spatiales très différentes," *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, 2006.

[70] J. Branke, K. Deb, K. Miettinen, and Slowinski, "Multiobjective Optimization," *Lectures Notes in Computer Sciences*, vol. 5252, pp. 1-470, 2008.

[71] Sytze de Bruin, "Modelling Positional Uncertainty of Line Features by Accounting for Stochastic Deviations from Straight Line Segments," *Transactions in GIS*, vol. 12, no. 2, pp. 165-177, 2008.

[72] P.A. Burrough and A.U. Frank, , Taylor \& Francis, Ed., 1987, ch. 12, pp. 171-187.

[73] J.R. Busby, *Biodiversity mapping and monitoring*, Skidmore Environmental Modelling with GIS and Remote Sensing A, Ed. London: Taylor and Francis, 2002.

[74] J. Cardoso, "The semantic web vision: Where are we?," *IEEE Intelligent Systems*, vol. 22, pp. 84–8, 2007.

[75] T. Carron, "Segmentation d'images couleur dans la base Teinte-Luminance-Saturation: approche numérique et symbolique," Université de Savoie, France, Ph.D. dissertation 2006.

[76] Yi lai Chen, TaoWang, Ben shengWang, and Zhou jun Li, "A Survey of Fuzzy Decision Tree Classifier," *Fuzzy Inf. Eng.*, vol. 2, pp. 149-159, 2009.

[77] Eliseo Clementini and Robert Laurini, "Un cadre conceptuel pour modéliser les relations spatiales," *Revue des Nouvelles Technologies de l'Information*, vol. 1, pp. 1-18, Novembre 2008.

[78] J.P. Cocquerez and S. Phillip, *Analyse d'images : Filtrage et segmentation.*: Masson, 1995.

[79] Olivier Colliot, Oscar Camara, and Isabelle Bloch, "Un modèle déformable intégrant des relations spatiales pour la segmentation de structures cérébrales," *Information interaction*

intelligence, vol. 5, no. 1, pp. 29-58, 2005.

[80] Guianni Commin, "Caractérisation de textures forestières," Rapport de stage de DEA, Université des Antilles et de la Guyane, Tech. rep. 2004.

[81] Coppock and Rhind, , Longman Scientific and Technical, Eds. London: D. J. Maguire and M. F. Goodchild and D. W. Rhind, 1991, ch. The History of GIS.

[82] Guilem Coq, Olivier Alata, Christian Olivier, and M. Arnaudon, "Méthodes comparatives d'utilisation des critères d'information pour la sélection de modèles," in *TAIMA*, Tunisie, mai 2009, pp. 241-246.

[83] S. Coros, JingBo Ni, and Pascal Matsakis, "Object Localization Based on Directional Information: Case of 2D Vector Data," in *14th Int. Symposium on Advances in Geographic Information Systems (ACM-GIS'06)*, 2006.

[84] Helen Couclesis, "People Manipulate Objects (but cultivate fields) : beyond the raster-Vector debate in GIS," in *Theories and Methods of Spatial-Temporal Reasoning in Geographic Space*, Springer, Ed.: A. U. Frank, I. Campari, and U. Formentini, 1992, pp. 105-138.

[85] V.V. Cross, "Fuzzy extensions for relationships in a generalized object model," *International Journal on Intelligent Systems*, vol. 16, pp. 843-861, 2001.

[86] I. Cruz, W. Sunna, and K Ayloo, "Concept-level matching of geospatial ontologies," *In Proceedings of GISPlanet*, 2005. [Online]. http://www.cs.uic.edu/~advis/William/finalpapers/FinalGISPlanet2005.pdf

[87] Nadine Cullot, Christine Parent, Stefano Spaccapietra, and Christelle Vangenot, "Des ontologies pour données géographiques," *Revue Internationale de Géomatique Numéro spécial "Les SIG sur le Web"*, vol. 13, no. 3, pp. 285-306, 2003.

[88] Dalle, Les contours déformables, 2010.

[89] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 182–197, 2002.

[90] Boriana Deliiska, "Thesaurus and Domain Ontology of Geoinformatics," *Transactions in GIS*, vol. 11, no. 4, pp. 637–651, 2007.

[91] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm (with discussion)," *Journal of the Royal Statistical Society*, vol. 39, 1979.

[92] C. Ding and H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," in *Second IEEE Computational Systems Bioinformatics Conf*, 2003, pp. 523-528.

[93] L.K.A. Dorren, B. Maier, and A.C. Seijmonsbergen, "Improved Landsat-based forest mapping in steep mountainous terrain using object-based classification," *Forest Ecology and Management*, vol. 183, pp. 31-46, 2003.

[94] M. J. Egenhofer and R. D. Franzosa, "Point-Set Topological Spatial Relations," *International Journal of Geographical Information Systems*, vol. 5, no. 2, pp. 161-174, 1991.

[95] Max J. Egenhofer and Jayant Sharma, "Assessing the consistency of complete and incomplete topological information," *Geographical Systems*, vol. 1, no. 1, pp. 47-68, 1993.

[96] Ikram El-Missi, "Application de l'outil fractal, multifractal, pour la séparation forêt / agriculture dans le cadre du projet PARAGE sur l'île de la Guadeloupe," Rapport de stage de fin d'étude, Master 2ème année, Université des Antilles et de la Guyane, Tech. rep. 2007.

[97] P. F. Fisher, *Classics from IJGIS: Twenty Years of the International Journal of Geographical Information Science*, Figsher, Ed.: CRC Hoboken, 2006.

[98] P. Fisher, "Sorites paradox and vague geographies," *Fuzzy Sets and Systems*, vol. 113, pp. 7-18, 2000.

[99] Larry Foisy, "Système d'extraction automatique de mesures forestières primaires à partir d'images satellites à très haute résolution spatiale," rapport de stage de M.Sc., Faculté des Sciences de l'Université de Sherbrooke, Canada, Tech. rep. 2007.

[100] J.D. Foley, A. Van Dam, S.K. Feiner, and J.F. Hughes, *Computer Graphics : Principles and Practice in C*, Addison-Wesley, Ed.: Longman, 1990.

[101] F. Fonseca, G. Camara, and A. M. Monteiro, "A framework for measuring the interoperability of geo-ontologies.," *Spatial Cognition and Computation*, vol. 6, pp. 309–31, 2006.

[102] F. Fonseca, M. Egenhofer, P. Agouris, and G. Câmara, "Using ontologies for integrated geographic information systems," *Transactions in GIS*, vol. 6, pp. 231–57, 2002.

[103] F. Fonseca, J. Martin, and A. Rodríguez, "From geo to eco-ontologies," *In Proceedings of the Second International Conference on Geographic Information Science*, pp. 93–107, 2002.

[104] Mark Gahegan, Ritesh Agrawal, Anuj Jaiswal, and Junyan Luo, "A Platform for Visualizing and Experimenting with Measures of Semantic Similarity in Ontologies and Concept Maps," *Transactions in GIS*, vol. 12, no. 6, pp. 713–732, 2008.

[105] Joao Gama, *Functional Trees*, Logistic Model Trees, Ed.: Niels Landwehr, Mark Hall, Eibe Frank, 2004.

[106] Christopher M Gold, "What is GIS and What is Not?," *Transactions in GIS*, vol. 10, no. 4, pp. 505-519, 2006.

[107] R.C. Gonzalez and R.E. Woods, *Digital Image Processing, 3rd ed.*: Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[108] R.C. Gonzalez and R.E. Woods, *Digital Image Processing, 3rd ed.*, Prentice-Hall Inc., Ed.: Upper Saddle River, 2006.

[109] Michael Frank Goodchild, "Twenty years of progress: GIScience in 2010," *Journal Of Spatial Information Science,* vol. 1, pp. 3-20, 2010.

[110] Michael F. Goodchild, Jingxiong Zhang, and Phaedon C. Kyriakidis, "Discriminant Models of Uncertainty in Nominal Fields," *Transactions in GIS*, vol. 13, no. 1, pp. 7-23, 2009.

[111] Daniel A. Griffith, Marco Millones, Matthew Vincent, David L. Johnson, and Andrew Hunt, "Impacts of Positional Error on Spatial Regression Analysis: A Case Study of Address Locations in Syracuse, New York," *Transactions in GIS*, vol. 11, no. 5, pp. 655-679, 2007.

[112] D. Guerin, F. Cointault, C. Gée, and J.P Guillemin, "Etude de faisabilité d'un système de comptage d'épis de blé par vision.," *Revue Traitement du Signal, Session Spéciale Imagerie Couleur*, vol. 21, no. 5, pp. 549–560, 2005.

[113] Hans W. Guesgen and Jochen Albrecht, "Imprecise reasoning in geographic information systems," *Fuzzy Sets and Systems*, vol. 113, pp. 121-131, 2000.

[114] Danni Guo, Renkuan Guo, and Christien Thiart, "Integrating GIS with Fuzzy Logic and Geostatistics: Predicting Air Pollutant PM10 for California, Using Fuzzy Kriging," Department of Statistical Sciences, University of Cape Town, Cap Town, Tech. rep. 2004.

[115] Filip Hájek, "Process-based approach to automated classification of forest structures using medium format digital aerial photos and ancillary GIS information," *Eur J Forest Res*, vol. 127, pp. 115–124, 2008.

[116] Marc Hallin and Abdelaziz El Matouat, "Order selection, stochastic complexity and Kullback-Leibler information," ULB -- Universite Libre de Bruxelles, Tech. rep. 1996. [Online]. http://ideas.repec.org/p/ulb/ulbeco/2013-2153.html

[117] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, and B.I. Shraiman, "Fractal measures and their singularities : The characterization of strange sets," *Physical Review*, vol. 33, no. 2, pp. 1141–1151, 1986.

[118] http://www.savgis.org/supports-de-cours.html,.

[119] M.K. Hu, "Visual pattern recognition by moment invariants," *Transactions on Information Theory*, vol. 8, no. 2, pp. 179-187, 1962.

[120] Céline Hudelot, Jamal Atif, and Isabelle Bloch, "Ontologie de relations spatiales floues pour le raisonnement spatial dans les images," in *RNTI*, 2006, pp. 55-86.

[121] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.

[122] Steve Jobs, You've got to find what you love - Stanford University Commencement address - http://news.stanford.edu/news/2005/june15/jobs-061505.html, June 2005.

[123] C.B. Jones and al, "Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project," *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 387-388, August 2002.

[124] Wolfgang Kainz, "Fuzzy Logic and GIS," Department of Geography and Regional Research, University of Vienna, Tech. rep. 2009.

[125] Wolfgang Kainz, "Introduction to Fuzzy Logic and Applications in GIS – Example," Department of Geography and Regional Research, University of Vienna, Tech. rep. 2009.

[126] T. Kanade, Y.I. Ohta, and T. Sakai, "Color information for region segmentation," *Computer Graphics and Image Processing*, vol. 13, pp. 222-241, 1980.

[127] M. Karimi, M.B. Menhaj, and M.S. Mesgari, "Preparing Mineral Potential Map Using Fuzzy Logic In GIS Environment," in *ISPRS*, vol. XXXVII, 2008.

[128] Anne Karsenty, Alzir Felippe, B. Antunes, and Jorge Silva Centeno, "Classification orientée objet de la perméabilité des sols en zone urbaine à l'aide d'imagerie très haute résolution et de données laser scanner à Curitiba (Brésil)," in *Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis,* 2007, pp. 565-572.

[129] M. Kavouras, M. Kokla, and E. Tomai, "Comparing categories among geographic ontologies," *Computers and Geosciences*, vol. 31, pp. 145–54, 2005.

[130] R. Kohavi and G. John, "Wrapper for Feature Subset Selection," *Artificial Intelligence*, vol. 1, pp. 273-324, 1997.

[131] K.I. Laws, "Rapid texture identification," in *SPIE Image Processing for Missile Guidance*, 1980, pp. 376–380.

[132] Sébastien Lefèvre and Nicole Vincent, "Apport de l'espace teinte-saturation-luminance pour la segmentation spatiale et temporelle," *Traitement du Signal*, vol. 23, no. 1, pp. 59–77, 2006.

[133] C. Lemaréchal, R. Fjørtoft, P. Marthon, and E. Cubero-Castan, "Comments on `Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 7, pp. 762-763, July 1998.

[134] H. Levkowitz and G.T. Herman, "Glhs : a generalized lightness, hue, and saturation color model," *CVGIP : Graphical Models and Images Processing*, vol. 55, no. 4, pp. 271–285, 1993.

[135] Stanislaw Lewinski and Karol Zaremski, "Examples of Object Oriented Classification Performed On High Resolution Satellite Images," *Miscellanea Geographica*, vol. 11, pp. 349-358, 2004.

[136] Jianning Liang, Su Yang, and Adam C. Winstanley, "Invariant optimal feature selection: A distance discriminant and feature ranking based solution," *Pattern Recognition*, vol. 41, no. 5, pp. 1429-1439, 2008.

[137] T.M. Lillesand, W.J. Carper, and R.W. Kiefer, "The use of Intensity-Hue-Saturation transformation for merging SPOT panchromatic and multispectral image data," *Photogrammetric Engineering and Remote Sensing*, vol. 56, 1990.

[138] E. Littmann and H. Ritter, "Adaptive color segmentation - a comparison of neural and statistical methods," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 175–185, 1997.

[139] J. Liu, "Smoothing filter-based intensity modulation : a spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.

[140] K. Liu and W. Shi, "Quantitative fuzzy topological relations of spatial objects by induced fuzzy topology," *International Journal of Applied Earth Observation and Geoinformation*, vol. 11, pp. 38-45, 2009.

[141] H. Liu, J. Sun, L. Liuand, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330-1339, 2009.

[142] D. L. Macadam, *Color Measurement, Theme and Variations, 2nd ed.*: Springer v, 1985.

[143] Ludovic Macaire, "Exploitation de la couleur pour la segmentation et l'analyse d'images," Université des sciences et technologies de Lille 1, Ph.D. dissertation 2004.

[144] Stephane G. Mallat, "A theory for multiresolution signal decomposition : the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11,

no. 7, pp. 674-693, 1989.

[145] Stephane G Mallat and H. Wang, "Singularity Detection And Processing With Wavelets," *Transactions on information theory*, vol. 38, no. 2, pp. 617-643, 1992.

[146] Stephane G Mallat and Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 14, no. 7, pp. 710-732, 1992.

[147] B. B. Mandelbrot, *Fractals : Form, Chance and Dimension.*: W. H. Freeman and Co, 1977.

[148] B.S. Manjunath and W.Y. MA, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 18, pp. 837–842, 1996.

[149] Laurent Manyri, "Développement des capteurs logiciels pour la caractérisation des populations microbiennes. Application aux procédés de fermentation : énergies renouvelable, boissons alcoolisées," Université des Antilles et de la Guyane, Ph.D. dissertation 2005.

[150] T. Mary-Huard and E. Lebarbier, "Une introduction au critère BIC : fondements théoriques et applications," *Journal de la société française de Statistiques*, vol. 147, no. 1, pp. 39-57, 2006.

[151] S. Mäs, "Reasoning on spatial relations between entity classes," in *Geographic Information Science, 5th International Conference, GIScience*, vol. 5266, 2008.

[152] S. Mäs, "Reasoning on spatial semantic integrity constraints," in *Spatial Information Theory, 8th International Conference, COSIT 2007*, vol. 4736, 2007.

[153] S. Mas, F. Wang, and W. Reinhardt, "Using ontologies for integrity constraint definition.," , 2005.

[154] Pascal Matsakis, , Pascal Matsakis and L. Sztandera, Eds.: Springer-Verlag, 2002, ch. Understanding the Spatial Organization of Image Regions by Means of Force Histograms: A Guided Tour, pp. 99-122.

[155] Pascal Matsakis, "Relations spatiales structurelles et interprétation d'images," Paul Sabatier University, Toulouse, France, Ph.D. dissertation 1998.

[156] Pascal Matsakis, Serge Andrefouet, and Patrick Capolsini, "Evaluation of Fuzzy Partitions," *Remote Sensing of Environment*, vol. 74, no. 3, pp. 516-533, 2000. [Online]. http://www.sciencedirect.com/science/article/pii/S0034425700001437

[157] Pascal Matsakis, J. Keller, and L. Wendling, "F-Histograms and Fuzzy Directional Spatial Relations," in *Conference on Fuzzy Logic and Its Applications (LFA)*, vol. 1, Valencienne, France, October 1999, pp. 207-213.

[158] Pascal Matsakis, J.M. Keller, L. Wendling, J. Marjamaa, and O. Sjahputera, "Linguistic description of relative positions in images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 4, pp. 573-588, 2001.

[159] Pascal Matsakis and D. Nikitenko,.: Springer-Verlag, 2005, ch. Combined Extraction of Directional and Topological Relationship Information from 2D Concave Objects, pp. 15-40.

[160] Pascal Matsakis, JingBo Ni, and M. Veltman, "Directional relationships to a reference object: A quantitative approach based on force fields," in *Proc. 16th IEEE Int Image Processing (ICIP) Conf*, 2009, pp. 321-324.

[161] Pascal Matsakis, JingBo Ni, and Xin Wang, "Object Localization Based on Directional Information: Case of 2D Raster Data," in *Proc. 18th Int. Conf. Pattern Recognition ICPR 2006*, vol. 2, 2006, pp. 142-146.

[162] Jennifer Miller, Janet Franklin, and Richard Aspinall, "Incorporating spatial dependence in predictive vegetation models," *Ecological Modelling*, vol. 20, no. 2, pp. 225–242, 2007.

[163] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge, Discovery, Second Edition*, Harvey J. Miller, Ed.: CRC Press, may 2009.

[164] Alina Dia Miron, Jérôme Gensel, Marlène Villanova-Oliver, and Hervé Martin, "Relations spatiales qualitatives dans les ontologies géographiques avec ONTOAST," *SAGEO*, 2007.

[165] B. Mukhopadhyay, Integrating exploration dataset in GIS using fuzzy inference modeling, 2002.

[166] Jingbo Ni and Pascal Matsakis, "An equivalent definition of the histogram of forces: Theoretical and algorithmic implications," *Pattern Recognition*, vol. 43, no. 4, pp. 1607-1617, 2010. [Online]. http://www.sciencedirect.com/science/article/pii/S0031320309003641

[167] JingBo Ni, M. Veltman, and Pascal Matsakis, "Directional Force Field-Based Maps: Implementation and Application," in *13th IAPR Int. Conf. on Computer Analysis of Images and Patterns (CAIP)*, Germany, Munster, september 2009, pp. 309-17.

[168] Y. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," *Computer Graphics Vision and Image Processing (CGVIP)*, vol. 13, pp. 222-241, 1980.

[169] M. Oruc, A. M. Marangoz, and G. Buyuksalih, "comparison of pixel-based and object-oriented classification approaches using Landsat-7 ETM spectral bands," in *Proceedings of the ISRPS 2004 Annual Conference*, 2004.

[170] J.N. Paoli, O. Strauss, B. Tisseyre, and P. Lagacherie, "Utilisation d'un variogramme flou dans une méthode d'agrégation sémantique," in *LFA*, 2006.

[171] G. Parisi and U. Frish, "Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics," *The Physical Society of Japan (JPS), North-Holland, Amsterdam and New York*, vol. 41, 1985.

[172] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *TPAMI*, vol. 27, no. 8, pp. 1226-1238, 2005.

[173] A. Porebski, N. Vandenbroucke, and L. Macaire, "Comparison of feature selection schemes for color texture classification," in *Int. Conf. on IPTA*, Paris, july 2010.

[174] Projet PARAGE (occuPation Agricole dans les Régions Antilles et Guyane). Spot Image, CIRAD, IRD, SIGBEA : http://parage.sigbea.fr (dernier accès Juin 2008).

[175] P. Pudil and J. Novovicova, "Novel methods for subset selection with respect to problem knowledge," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 66-74, 1998.

[176] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1199-1207, 2005.

[177] Ross Quinlan, *C4.5: Programs for Machine Learning.*, San Mateo CA, Ed.: Morgan Kaufmann, 1993.

[178] G. Raines and al., "New fuzzy logic tools in ArcGIS 10," ESRI Communication, Tech. rep. 2010.

[179] Robert G. Raskin and Michael J. Pan, "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)," *Computers & Geosciences*, vol. 31, pp. 1119–1125, 2005.

[180] Amandine Robin and Sylvie Le Hégarat-Mascle, "An a-contrario approach for sub-pixel change detection in satellite imagery'','' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1977-1993, 2010.

[181] M.A. Rodriguez and M.J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 442-56, 2003.

[182] Alain Rousteau, "Carte écologique de la Guadeloupe. 3 feuilles au 1/75.000ème et notice (36p)," Conseil Général de la Guadeloupe, Office National des Forêts et Parc National de la Guadeloupe, Tech. rep. 1996.

[183] M.C. Ruiz and al., "The Development of a New Methodology Based on GIS and Fuzzy Logic to Locate Sustainable Industrial Areas," in *10th AGILE International Conference on Geographic Information Science*, 2007.

[184] Y. Sakamoto and H. Akaike, "Analysis of cross classified data by AIC," *Ann. Inst. Statist. Math*, vol. 30, no. B, pp. 185-197, 1978.

[185] D. Sawatzky, G. Raines, and G. Bonham-Carter, "Spatial Data Modeller," Arc SDM, Tech. rep. 1999.

[186] Markus Schneider, *Spatial Data Types for Database Systems Finite Resolution Geometry for Geographic Information Systems.*: Lecture Notes in Computer Science, 1997, vol. 1288.

[187] Markus Schneider, "Uncertainty management for spatial data in databases: fuzzy spatial data types," *Advances in Spatial Databases, Lecture Notes in Computer Science*, vol. 1651, pp. 330-351, 1999.

[188] Nadine Schuurman and Agnieszka Leszczynski, "Ontology-Based Metadata," *Transactions in GIS*, vol. 10, no. 5, pp. 709-726, 2006.

[189] H.C. Shyu, T.M. Tu, S.C Su, and P.S Huang, "A new look at IHS-like image fusion methods," *Information Fusion*, vol. 2, 2001.

[190] S.C. Sides, P.S. Chavez, and J.A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT Panchromatic," *Photogrammetric Engineering \& Remote Sensing*, vol. 57, 1991.

[191] SIG La Lettre n°111, novembre 2009.

[192] Philip Smart, Alia Abdelmoty, Baher El-Geresy, and Christopher Jones, "A Framework for Combining Rules and Geo-ontologies," in *Web Reasoning and Rule Systems* , Massimo Marchiori, Jeff Z. Pan, and Christian Marie, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4524, ch. 10, pp. 133-147. [Online]. http://dx.doi.org/10.1007/978-3-540-72982-2_10

[193] R. Sunila, "Fuzzy modelling and Kriging for modelling imprecise soil polygon boundaries," in *International Conference on Geoinformatics-Geospatial Information Research: Bridging the Pacific and Atlantic*, 2004, pp. 489-495.

[194] R. Sunila and P. Horttanainen, "Fuzzy Model of Soil Polygons for Managing the Imprecision," *Interfacing GeoStatistics and GIS*, 2009.

[195] Y. Sun, S. Todorovic, and S. Goodison, "Local-Learning-Based Feature Selection for High-Dimensional Data Analysis," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 32, no. 9, pp. 1610-1626, 2010.

[196] Francesco Tarquini and Eliseo Clementini, "Spatial Relations between Classes as Integrity Constraints," *Transactions in GIS*, vol. 12, pp. 45–57, 2008.

[197] Alban Thomas, "Application de l'approche orientée-objet à l'extraction de fragments forestiers à partir de scènes Spot," Rapport de stage, DESS SIGMA, Tech. rep. 2005.

[198] C. Thomas, "Fusion d'images de résolutions spatiales différentes," L'Ecole des Mines de Paris, Ph.D. dissertation 2006.

[199] D. Travis, *Effective colour displays : Theory and Practice 1ed ed.*: Academic Press, 1991.

[200] M. Unser, "Sum and difference histograms for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 8, no. 1, pp. 118-125, 1986.

[201] N. Vandenbroucke, "Segmentation d'images couleur par classification de pixels dans des espaces d'attributs colorimétriques adaptés. Application à l'analyse d'image de football," Université des Sciences et Technologies de Lille 1, Ph.D. dissertation 2000.

[202] N. Vandenbroucke, L. Macaire, and J.G. Postaire, "Color image segmentation by pixel classification in an adapted hybrid color space. application to soccer image analysis," *Computer Vision and Image Understanding*, vol. 90, no. 2, pp. 190–216, 2003.

[203] V. Vapnik, *The nature of statistical learning theory, 2nd ed.*: Springer verlag, 1999.

[204] H. Veregin, "Data quality parameters," *Geographical Information Systems Principles and Technical Issues*, pp. 177–189, 1999.

[205] R. Viegas and V. Soares, "Querying a Geographic Database using an Ontology-Based Methodology," in *GEOINFO 2006 - VIII Brazilian Symposium on GeoInformatics*, Brazil, 2006.

[206] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *3th conference: Fusion of Earth data: merging point measurements, raster maps and remotely sensed images*, 2000.

[207] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions : assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[208] X. Wang, P. Matsakis, L. Trick, B. Nonnecke, and M. Veltman, "A Study on How Humans Describe Relative Positions of Image Objects," in *SDH 2008 (13th Int. Symposium on Spatial Data Handling)*, Montpellier, France, June 2008, pp. 1-18.

[209] Xin Wang, JingBo Ni, and Pascal Matsakis, "Fuzzy Object Localization Based on Directional (and Distance) Information," in *Proc. IEEE Int Fuzzy Systems Conf*, 2006, pp. 256-263.

[210] Lukasz Wawrzyniak, Pascal Matsakis, and Dennis Nikitenko, "Representing Topological

Relationships between Complex Regions by F-Histograms," *Int. J. Intelligent Systems Technologies and Applications*, vol. 1, no. 3/4, pp. 1-12, 2006.

[211] Lukasz Wawrzyniak, Dennis Nikitenko, and Pascal Matsakis, "Speaking with spatial relations," *Int. J. Intelligent Systems Technologies and Applications*, vol. 1, no. 3/4, pp. 280-300, 2006.

[212] T. H. Wong, S. B. Mansor, M. R. Mispan, N. Ahmad, and W.N.A. Sulaiman, "Feature extraction based on object oriented analysis," in *ATC*, Malesia, May 2003.

[213] D.J. Wright, Michael Frank Goodchild, and J.D. Proctor, "Demystifying the persistent ambiguity of GIS as 'tool' versus 'science'," *Annals of the Association of Americn Geographers*, vol. 87, no. 2, pp. 346-362, 1997.

[214] T.A. Yanar and Z. Akyürek, "The Enhancement of ArcGIS with Fuzzy Set Theory," in *ESRI International User Conference*, 2004.

[215] Qian Yu et al., "Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery," *Photogrammetric Engineering \& Remote Sensing*, vol. 72, no. 7, pp. 799–811, July 2006.

[216] C. Zaki, M. Servières, and Guillaume Moreau, "Implementing conceptual spatiotemporal model into object dbms with semantic preserving," in *IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, Fuzhou, China, July 2011.

[217] C. Zaki, M. Servières, and Guillaume Moreau, "Transforming conceptual spatiotemporal model into Object model with semantic keeping," in *5th International workshop on Semantic and Conceptual Issues in GIS (SeCoGis 2011)*, Brussels, Belgium, 2011.

[218] Paul A Zandbergen, "Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National standard for Spatial Data Accuracy," *Transactions in GIS*, vol. 12, no. 1, pp. 103–130, 2008.

[219] A. X. Zhu and al., "Soil Mapping Using GIS," *Expert Knowledge, and Fuzzy Logic*, vol. 65, pp. 1463–1472, 2001.