



Apprentissage automatique des classes d'occupation du sol et représentation en mots visuels des images satellitaires

Marie Lauginie Lienou

► To cite this version:

Marie Lauginie Lienou. Apprentissage automatique des classes d'occupation du sol et représentation en mots visuels des images satellitaires. Traitement du signal et de l'image. Télécom ParisTech, 2009. Français. <[pastel-00005585](#)>

HAL Id: [pastel-00005585](#)

<https://pastel.archives-ouvertes.fr/pastel-00005585>

Submitted on 19 Oct 2010

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

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



Thèse

Présenté pour obtenir le grade de docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Signal et Images**

Marie Lauginie LIENOU

APPRENTISSAGE AUTOMATIQUE DES CLASSES D'OCCUPATION DU
SOL ET REPRÉSENTATION EN MOTS VISUELS DES IMAGES
SATELLITAIRES

soutenue le 02-03-2009 devant le jury composé de :

Patrick Gros
Philippe Bolon
Jocelyn Chanussot
Harald Kosch
Jordi Inglada
Henri Maître
Mihai Datcu

Président
Rapporteurs
Examinateurs
Directeurs de thèse

"Le doute est à la base même du savoir, puisqu'il est la condition essentielle de la recherche de la vérité. On ne court jamais après ce que l'on croit posséder avec certitude."

Jean-Charles Harvey, *Les demi-civilisés.*

*To my beloved parents,
To my sister and brothers.*

Remerciements

Comme dit le proverbe, “la reconnaissance silencieuse ne sert à personne”. Durant ma thèse, nombreux sont ceux qui m’ont apporté une contribution scientifique, logistique ou morale. Je tiens donc à leur exprimer toute ma gratitude.

J’aimerais tout d’abord remercier mon directeur de thèse, Henri Maître, qui a toujours veillé à ce que ma thèse se déroule dans de bonnes conditions de travail, et dont l’aide précieuse m’a été indispensable tant sur le plan scientifique qu’humain. Je remercie également Mihai Datcu qui a co-dirigé ma thèse, pour ses idées, ses petits dessins explicatifs, ainsi que sa disponibilité.

Je tiens à exprimer ma profonde gratitude à Patrick Gros, qui m’a fait l’honneur de présider mon jury de thèse. Je suis très reconnaissante à Philippe Bolon et Jocelyn Channusot pour avoir accepté le rôle de rapporteurs. Leurs commentaires et questions ont fortement contribué à améliorer ce document. Je remercie également Harald Kosch et Jordi Inglada, pour avoir accepté d’examiner mon mémoire et pour l’intérêt qu’ils ont porté à mon travail.

Je ne saurais ne pas témoigner ma reconnaissance à Marine Campedel, dont l’aide sur le plan technique et les grandes qualités humaines m’ont permis de mener à bien cette thèse. Son dynamisme au sein du Centre de Compétences, et son optimisme ont très souvent été contagieux. Je tiens aussi à la remercier pour la confiance et la sympathie qu’elle m’a témoignées au cours de ces années de thèse et de master.

Je remercie également Alain Giros, pour les discussions très constructives, sa relecture de mon manuscrit et ses conseils pertinents.

Ma gratitude s’adresse aussi aux permanents du département TSI, en particulier à Patricia et Isabelle, et à mes collègues de bureau avec qui j’ai passé d’excellents moments : merci à Gabrielle pour sa solidarité féminine, Mihai pour les bonnes blagues, Thibault pour m’avoir fait découvrir BattleStar Galactica, Julien pour ses anecdotes et ses équations au tableau, Thomas pour ses “jour J moins ...”, Ivan pour sa disponibilité, sans oublier Camille, Jérémie, Aymen, Charles et les autres.

Je tiens en outre à remercier Patrice pour son soutien indéfectible et son enthousiasme manifeste à l’égard de mes travaux. Sa détermination, communicative, et ses expressions avisées telles que “Aucun effort n’est vain !”, ont été de véritables stimulants pour ma motivation.

“Last but not least”, je remercie les parents et amis qui, dans l’ombre, m’ont toujours encouragé. Leur soutien s’est avéré déterminant pour mener ce travail à terme. Merci encore !

Résumé

La reconnaissance de la couverture des sols à partir de classifications automatiques est l'une des recherches méthodologiques importantes en télédétection. Par ailleurs, l'obtention de résultats fidèles aux attentes des utilisateurs nécessite d'aborder la classification d'un point de vue sémantique. Cette thèse s'inscrit dans ce contexte, et vise l'élaboration de méthodes automatiques capables d'apprendre des classes sémantiques définies par des experts de la production des cartes d'occupation du sol, et d'annoter automatiquement de nouvelles images à l'aide de cette classification.

A partir des cartes issues de la classification CORINE Land Cover, et des images satellitaires multispectrales ayant contribué à la constitution de ces cartes, nous montrons tout d'abord que si les approches classiques de la littérature basées sur le pixel ou la région sont suffisantes pour identifier les classes homogènes d'occupation du sol telles que les champs, elles peinent cependant à retrouver les classes de haut-niveau sémantique, dites de mélange, parce qu'êtants composées de différents types de couverture des terres.

Pour détecter de telles classes complexes, nous représentons les images sous une forme particulière basée sur les régions ou objets. Cette représentation de l'image, dite en mots visuels, permet d'exploiter des outils de l'analyse de textes qui ont montré leur efficacité dans le domaine de la fouille de données textuelles et en classification d'images multimédia. A l'aide d'approches supervisées et non supervisées, nous exploitons d'une part, la notion de compositionnalité sémantique, en mettant en évidence l'importance des relations spatiales entre les mots visuels dans la détermination des classes de haut-niveau sémantique. D'autre part, nous proposons une méthode d'annotation utilisant un modèle d'analyse statistique de textes : l'Allocation Dirichlet Latente. Nous nous basons sur ce modèle de mélange, qui requiert une représentation de l'image dite en sacs-de-mots visuels, pour modéliser judicieusement les classes riches en sémantique. Les évaluations des approches proposées et des études comparatives menées avec les modèles gaussiens et dérivés, ainsi qu'avec le classificateur SVM, sont illustrées sur des images SPOT et QuickBird entre autres.

Mots-clés : apprentissage automatique, modèle LDA, classification, sémantique, représentation en mots visuels, images satellitaires, cartographie.

Abstract

Land cover recognition from automatic classifications is one of the important methodological researches in remote sensing. Besides, getting results corresponding to the user expectations requires approaching the classification from a semantic point of view. Within this frame, this work aims at the elaboration of automatic methods capable of learning classes defined by cartography experts, and of automatically annotating unknown images based on this classification.

Using CORINE Land Cover maps, we first show that classical approaches in the state-of-the-art are able to well-identify homogeneous classes such as fields, but have difficulty in finding high-level semantic classes, also called mixed classes because they consist of various land cover categories.

To detect such classes, we represent images into visual words, in order to use text analysis tools which showed their efficiency in the field of text mining. By means of supervised and not supervised approaches on one hand, we exploit the notion of semantic compositionality : image structures which are considered as mixtures of land cover types, are detected by bringing out the importance of spatial relations between the visual words. On the other hand, we propose a semantic annotation method using a statistical text analysis model : Latent Dirichlet Allocation. We rely on this mixture model, which requires a bags-of-words representation of images, to properly model high-level semantic classes. The proposed approach and the comparative studies with Gaussian and GMM models, as well as SVM classifier, are assessed using SPOT and QuickBird images among others.

Keywords : machine learning, LDA model, classification, semantic level, visual words representation, satellite images, cartography.

Table des matières

1	Introduction	15
1.1	Contexte et problématique	15
1.2	Objectifs et contributions	16
1.3	Organisation du document	19
2	Apprentissage automatique et classification supervisée	21
2.1	Notion d'apprentissage automatique	21
2.1.1	Fouille de données	22
2.1.2	Catégories d'apprentissage automatique	23
2.2	Application à la classification supervisée	24
2.2.1	Principe	24
2.2.2	Etat de l'art en imagerie satellitaire	25
2.2.2.1	Algorithmes d'apprentissage	26
2.2.2.2	Types d'information	28
2.3	Conclusions	29
3	Processus de classification automatique supervisée	31
3.1	Du pixel à la région : la segmentation	31
3.2	Extraction de primitives	34
3.2.1	Primitives spectrales	34
3.2.2	Primitives de texture	36
3.2.3	Primitives géométriques	40
3.3	Sélection de primitives	41
3.4	Méthodes de classification supervisée	42
3.4.1	k-plus-proches voisins (k -ppv)	42
3.4.2	Classification bayésienne	43
3.4.2.1	Théorie de la décision bayésienne	43
3.4.2.2	Modèle bayésien naïf	45
3.4.2.3	Modèle de mélange de gaussiennes	46
3.4.3	Machines à vecteurs de support (SVM)	48
3.5	Evaluation de la classification	50
3.6	Sélection de modèles	51
3.6.1	Validation croisée	51
3.6.2	Les méthodes statistiques	52
3.6.2.1	Akaike Information Criterion	53
3.6.2.2	Bayesian Information Criterion	53
3.6.2.3	Minimum Description Length	53
3.7	Méthodes de régularisation	54

3.7.1	Vote majoritaire	54
3.7.2	Champs de Markov	55
3.8	Conclusions	56
4	Apprentissage automatique des classes d'occupation du sol	57
4.1	Description des données	57
4.1.1	CORINE Land Cover	58
4.1.2	Images SPOT	59
4.1.3	Gestion des données	60
4.1.4	Description de la zone de test	62
4.2	Méthodologie d'apprentissage et de classification	63
4.2.1	Protocole	63
4.2.2	Hypothèses	64
4.2.3	Combinatoire des choix	65
4.3	Classification sur critères radiométriques	67
4.3.1	Analyse des résultats	71
4.3.2	Comparaison de classifications	72
4.3.3	Traitement contextuel	72
4.3.4	Remarques	75
4.3.5	Solutions possibles	76
4.4	Prise en compte des textures	77
4.4.1	Caractéristiques QMF	77
4.4.2	Utilisation combinée des caractéristiques spectrales et texturelles	77
4.5	Amélioration de la classification	79
4.5.1	Classification hiérarchique	79
4.5.2	Regroupement des classes	82
4.5.3	Graphes d'adjacence	84
4.6	Applications sur un terrain inconnu	85
4.7	Conclusions	86
5	Représentation des images basée sur une approche par régions	89
5.1	Potentiel de l'approche par régions	89
5.1.1	Principe de compositionnalité sémantique	90
5.1.2	Approche par sacs-de-mots	91
5.2	Codage de l'image : les mots visuels	91
5.2.1	Le vocabulaire visuel	91
5.2.2	Etat de l'art	92
5.2.3	Approche choisie	93
5.3	Relations spatiales et représentation	94
5.3.1	Analyse syntaxique	94
5.3.2	Représentation des relations spatiales	95
5.4	Apprentissage	96
5.4.1	Annotation automatique d'images	97
5.4.1.1	Annotation directe	97
5.4.1.2	Annotation utilisant un niveau intermédiaire	98
5.4.2	Modèles de mélange de l'analyse de texte	101
5.4.2.1	Analyse sémantique latente probabiliste (pLSA)	101
5.4.2.2	Allocation Dirichlet Latente (LDA)	102

5.4.2.3	Utilisation pour la modélisation des images	104
5.5	Conclusions	107
6	Relations spatiales entre les mots visuels	109
6.1	Représentation des informations extraites	109
6.1.1	Les graphes d'adjacence	110
6.1.2	Les hypergraphes	110
6.1.3	Représentation de l'image en un texte	112
6.1.4	Analyse et choix d'une représentation	112
6.2	Détermination non supervisée de structures d'intérêt	115
6.2.1	Protocole	115
6.2.2	Détection des bâtiments et de leurs ombres	116
6.3	Apprentissage supervisé des relations entre les mots visuels	117
6.3.1	Méthodologie	118
6.3.2	Expérimentations sur les <i>forêts mélangées</i>	119
6.4	Conclusions	123
7	Annotation sémantique des images satellitaires basée sur le modèle LDA	125
7.1	Approche d'annotation sémantique de grandes images	125
7.1.1	Classification supervisée des imagettes	126
7.1.2	Prise en compte de l'information spatiale	128
7.2	Expérimentations et résultats	128
7.2.1	Influence de l'information spatiale	129
7.2.1.1	Description des données	129
7.2.1.2	Réglage des paramètres et tests	129
7.2.1.3	Analyse des résultats	131
7.2.2	Utilisation des caractéristiques radiométriques et texturelles	133
7.2.2.1	Description des données	133
7.2.2.2	Protocole de tests	133
7.2.2.3	Evaluation	134
7.2.2.4	Analyse des résultats	135
7.2.3	Cas de l'assimilation des mots à des régions de l'image	141
7.3	Etudes comparatives	142
7.3.1	Modèles gaussien, GMM et LDA	143
7.3.1.1	Expérimentations	143
7.3.1.2	Modèle gaussien	145
7.3.1.3	Modèles de mélange	148
7.3.1.4	Sélection de modèles	150
7.3.2	Classifications basées sur le LDA et le SVM	150
7.3.2.1	Classification SVM basée sur les sacs-de-mots	151
7.3.2.2	Classification SVM basée sur le pixel	153
7.4	Conclusions	154
8	Conclusions et perspectives	155
8.1	Conclusions	155
8.2	Perspectives	156

A CORINE Land Cover	161
A.1 Description	161
A.2 Nomenclature standard complète	164
A.3 Matrices de confusion	166
A.4 Classification : données issues de scènes différentes	166
B Descripteurs statistiques de Haralick	175
Bibliographie	177

Chapitre 1

Introduction

1.1 Contexte et problématique

Durant les dernières années, les progrès technologiques en termes d'acquisition des images (microscopes, caméras, capteurs) ont entraîné une augmentation considérable du nombre d'images disponibles. Par exemple, dans le domaine de la télédétection pour l'observation de la Terre, de grandes quantités d'images ont été délivrées par différents capteurs radars à synthèse d'ouverture (*Synthetic Aperture Radar* ou SAR), et optiques de type SPOT¹ ou Landsat². Plusieurs teraoctets sont encore produits chaque jour, grâce à une nouvelle génération de capteurs à haute résolution qui fournit en permanence de nouvelles images de la Terre. Le satellite QuickBird par exemple, est capable d'acquérir annuellement des données couvrant plus de 75 millions de km^2 . Par ailleurs, de nombreux autres satellites sont attendus dans les années à venir, notamment la famille de satellites optiques Pléiades de résolution sub-métrique. Ce large éventail d'images variées d'observation de la Terre est exploité dans divers domaines d'applications liés aux ressources naturelles et aux activités humaines : la reconnaissance des cultures, les risques environnementaux, la surveillance planétaire, l'aménagement du territoire et en particulier l'occupation des sols.

Par ailleurs, ce volume important d'images disponibles et les avancées en termes de stockage de données engendrent des bases de données de plus en plus riches en information. Il devient alors difficile pour l'humain, d'exploiter les informations contenues dans ces grandes bases d'images. En effet, les tâches de classification et d'indexation d'images effectuées manuellement nécessitent des efforts humains et financiers importants, le temps requis, le caractère répétitif de la tâche et la concentration nécessaire étant problématiques. Les cartes CORINE Land Cover³, utilisées à l'échelle européenne, illustrent bien cette situation. Ces cartes sont générées à partir d'images satellitaires qui sont interprétées visuellement par un expert, s'aidant de données exogènes (photographies aériennes, cartes topographiques et thématiques, etc). Des centaines de milliers de km^2 devant être décrits, le processus est long et délicat. Il serait donc intéressant de disposer de méthodes automatiques pour l'analyse et l'interprétation de telles bases d'images.

¹Satellite Pour l'observation de la Terre (<http://www.spotimage.fr>)

²Ce programme américain de télédétection spatiale (NASA) a été le premier programme civil d'observation de la Terre par satellite. Il a commencé avec le lancement du premier LANDSAT en juillet 1972.

³Le projet européen CORINE Land Cover a pour objectif de produire un inventaire biophysique de l'occupation des sols européens. Cette description repose sur une taxinomie précise afin d'assurer la consistance des résultats produits par les photointerprétes.

Le besoin d'automatiser les méthodes de classification et d'indexation se justifie en outre par la nécessité d'avoir des systèmes efficaces de gestion, d'accès et de recherche dans les banques d'images. Des études récentes menées à ce propos utilisent le contenu visuel des images pour les répertorier, c'est-à-dire qu'elles font usage de l'information portée par des attributs numériques directement calculés sur l'image, tels que la couleur ou la texture, encore appelés caractéristiques de "bas-niveau". Cependant, ces méthodes d'indexation par le contenu visuel ne sont pas suffisantes pour communiquer des résultats fidèles aux attentes des utilisateurs humains, car ces derniers effectuent plutôt des requêtes sémantiques par mots-clés ou en langage naturel (montagne, forêt, zone urbaine, etc). Ce problème se traduit par le "fossé sémantique" (ou *semantic gap*), qui est défini dans [Smeulders et al., 2000], comme "le manque de concordance entre les informations qu'on peut extraire des données visuelles et l'interprétation de ces mêmes données par un utilisateur dans une situation déterminée". Un accès efficace aux images, à un niveau sémantique, est donc un défi à relever pour permettre aux utilisateurs de profiter du contenu des bases d'images et répondre à leur besoin d'information.

Ainsi, l'organisation des bases de données en classes sémantiques dépendant du contexte de l'utilisateur est souhaitable, voire nécessaire. Dans ce contexte, nous nous intéressons dans cette thèse, à un problème actuel présent dans les tâches de classification et extraction du contenu des images satellitaires, à savoir la construction d'un système qui apprend de manière automatique, des informations de haut-niveau sémantique (c'est-à-dire définies par des utilisateurs) contenues dans une image.

1.2 Objectifs et contributions

La nécessité de disposer de méthodes d'apprentissage automatique pour la classification des images satellitaires présente un intérêt particulier pour les applications de type cartographie de l'occupation du sol. Les images satellitaires se prêtent particulièrement bien à la planification urbaine et rurale, puisque grâce à leur haute fréquence, elles donnent une image complète de la trame urbaine et de son arrière-pays. Elles forment une source de données fonctionnelles, facilement utilisables par les Systèmes d'Information Géographiques (SIG). Quelques exemples d'images satellitaires délivrées par différents capteurs optiques sont exposés dans la figure 1.1. Suivant la résolution des images, il est possible d'y identifier différents types d'occupation du sol, correspondant à des classes sémantiques différentes. Par exemple, dans les images à faible résolution telles que SPOT2 (figure 1.1(a)), on trouve des forêts, des champs, des villes, etc. Au fur et à mesure que la résolution croît, les classes identifiées sont plus localisées et plus riches en sémantique. Il s'agit par exemple des zones résidentielles, des routes, des bâtiments, des parkings, etc. Ainsi, suivant la résolution des images, les types d'occupation du sol recherchés par les experts dans les images diffèrent.

Pour de telles applications, nous souhaitons élaborer des méthodes capables d'apprendre automatiquement une taxinomie définie par des experts. La classification hiérarchique de CORINE Land Cover par exemple, repose sur une telle taxinomie et les cartes établies par les experts constituent une référence. Il serait donc intéressant d'utiliser cette classification en vue d'apprendre les règles de décision pour la classification automatique de terrains inconnus à partir d'images satellitaires.

Un large éventail de méthodes de classification ont été développées dans la littérature pour extraire des informations de couverture des sols (*land cover*) à partir d'images de

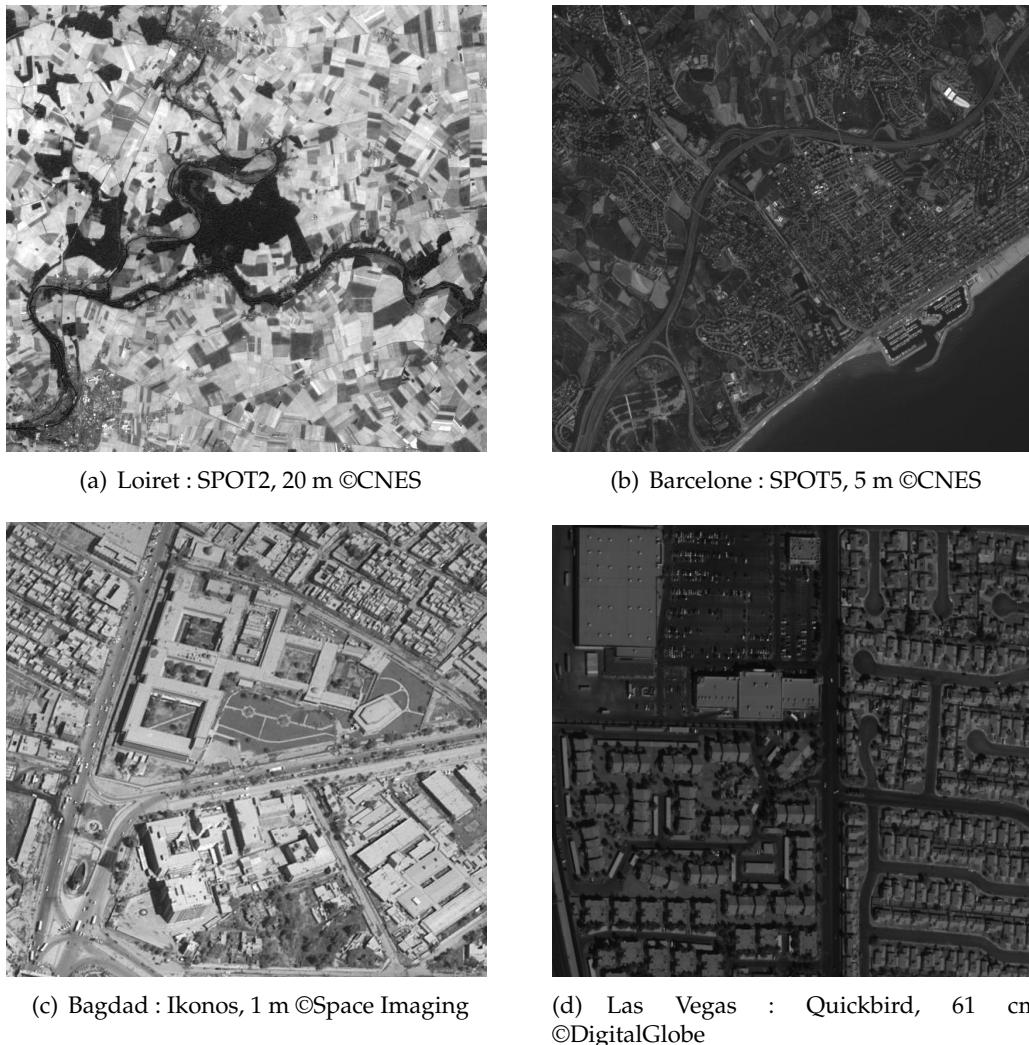


FIG. 1.1 – Exemples d’images délivrées par différents capteurs optiques, à différentes résolutions.

télédétection pour l’observation de la Terre. Une image étant un tableau de pixels, les méthodes classiques utilisent une approche basée sur le pixel, supervisée ou non [Jensen, 1986; Gong et al., 1992; Casals-Carrasco et al., 2000]. Chaque pixel de l’image est attribué à une classe, principalement à partir de propriétés locales (spectrales et/ou texturelles). La classification basée sur le Maximum de vraisemblance, qui est la méthode basée sur le pixel la plus utilisée, est limitée dans la mesure où elle ne tient pas compte de l’information contextuelle [Zhou & Robson, 2001; Dean & Smith, 2003]. Cette limitation se traduit par une mauvaise identification de certaines classes dites “de mélange”. En effet, il s’agit de classes qui sont composées de différents types de couverture des terres, et qui présentent des difficultés à être bien identifiées de manière automatique, avec de telles approches. Cette critique reste valable pour les approches orientées régions ou objets [Baraldi & Parmiggiani, 1990; Kusaka et al., 1990], qui considèrent l’image comme un ensemble de régions significatives obtenues via une segmentation de l’image effectuée au

préalable, et classifient chaque région indépendamment.

Dans cette thèse, les études que nous avons tout d'abord menées avec la taxinomie de CORINE Land Cover ont confirmé ces observations, notamment avec la classe des *forêts mélangées*, qui sont une combinaison de *forêts de feuillus* et de *forêts de conifères*, ces dernières étant elles-mêmes des classes à part entière. De même, les classes présentes dans des images à plus haute résolution telles que les *banlieues résidentielles* sont composées de plusieurs classes élémentaires (maisons, jardins, piscines, etc).

L'approche développée dans cette thèse pour pallier de telles difficultés repose sur le choix d'une représentation des images, motivé par les limitations rencontrées. Cette représentation de l'image, dite en mots visuels, est basée sur une approche orientée région. Mais contrairement aux méthodes de classification qui classifient chaque région, nous nous intéressons à une telle représentation pour ses capacités de mise en évidence des structures de l'image, facilitant ainsi son exploration et son interprétation. En outre, elle constitue une représentation concise de l'image, qui peut ensuite être utilisée à plusieurs fins, notamment avec les méthodes inspirées de l'analyse de textes, telles que les modèles à variable latente, qui ont déjà prouvé leur efficacité.

Partant de cette représentation, nous proposons d'exploiter le principe de compositionnalité sémantique [Pelletier, 1994], très utilisé en Traitement Automatique des Langues (TAL), et de l'appliquer aux images satellitaires afin d'analyser leur contenu, et en particulier identifier les classes de mélange. Ce principe établit que la signification d'une expression complexe est fonction de la signification de ses parties et de leurs relations syntaxiques. Nous proposons donc de mener l'opération de reconnaissance des classes de mélange, en identifiant des structures de l'image potentiellement porteuses de sémantique. Une structure étant considérée comme un groupe de mots visuels (régions élémentaires) connexes, la mise en évidence des structures est donc effectuée via la prise en compte des relations spatiales entre les mots visuels de l'image, ainsi que leur modélisation.

Par ailleurs, nous proposons une méthode d'annotation des grandes images satellitaires, basée sur un modèle d'analyse statistique de texte : les portions de l'image, qui dans notre cas peuvent être considérées comme des structures, sont associées à des concepts sémantiques définis au préalable par l'utilisateur. Il s'agit en fait d'une annotation sémantique de l'image abordée comme un problème de classification supervisée, les classes pouvant être pures ou de mélange. L'idée est de profiter de la hiérarchie du modèle textuel choisi, et du fait que la représentation de l'image que nous avons adoptée est appropriée pour ce modèle, pour mieux classifier les classes de mélange. En effet, ces dernières peuvent être vues comme une hiérarchie sémantique dans laquelle les composantes du mélange, qui sont les mots visuels dans le cas des structures, appartiendraient au niveau inférieur à celui de la classe de mélange.

Nous menons une étude comparative approfondie, théorique et expérimentale, entre la classification proposée et d'autres approches basées sur le modèle gaussien et le classificateur SVM (Séparateur à Vaste Marge ou Support Vector Machine). Il s'agit d'évaluer la contribution de l'approche par structures que nous proposons par rapport aux approches classiques, dans la détermination des classes de mélange en particulier.

1.3 Organisation du document

Pour présenter plus en détail nos réflexions sur les méthodes de classification automatique des images satellitaires pour les applications de type cartographie de l'occupation du sol, le présent document est structuré en huit chapitres.

Le chapitre 2 introduit la notion d'apprentissage automatique en général, et se focalise sur son application à la classification supervisée. Nous nous intéressons à l'imagerie de télédétection, et présentons un état de l'art des méthodes de classification d'images satellitaires pour les applications de cartographie de l'occupation du sol.

Dans le chapitre 3, nous détaillons chacune des étapes importantes du processus de classification supervisée des images basée sur un algorithme d'apprentissage automatique. Les différents algorithmes et outils de la classification supervisée y sont exposés, ainsi que les avantages et inconvénients de chacun d'eux. En effet, du choix des primitives, de la méthode de classification, de l'algorithme d'apprentissage automatique et éventuellement du modèle utilisé, dépendent les performances de la classification.

Le chapitre 4 quant à lui, illustre une application de l'apprentissage automatique pour la classification supervisée, tel qu'on peut le faire avec des approches classiques. En effet, les images satellitaires utilisées pour produire les cartes CORINE Land Cover sont analysées via une approche pixel, pour apprendre les classes sémantiques définies par des experts. La difficulté d'apprendre les classes de mélange et de les reconnaître dans de nouvelles images nous a conduit, dans le chapitre 5 à faire un choix de représentation de l'image pouvant permettre d'aborder ce problème.

En effet, la représentation en mots visuels est introduite dans le chapitre 5. Nous y montrons le potentiel d'une telle représentation et son intérêt pour notre problématique. Elle permet entre autres de pouvoir identifier les structures de l'image, tout en exploitant les outils de l'analyse de textes qui paraissent performants.

Le chapitre 6 exploite la compositionnalité sémantique et utilise les relations spatiales entre les mots visuels composant la représentation de l'image, pour déterminer les structures et les classes de mélange. Le chapitre 7 quant à lui, propose une méthode d'annotation de grandes images à l'aide d'un modèle de fouille de textes adapté à la représentation en mots visuels : l'Allocation Dirichlet Latente (LDA), utilisant une représentation dite en sacs-de-mots. Les classes sémantiques utilisées étant pures ou mélangées, la structure hiérarchique du modèle est exploitée pour la bonne identification de ces dernières classes. Enfin, nous comparons notre approche avec celles de la littérature, notamment les classificateurs utilisant les modèles gaussiens et mélanges de gaussiennes, mais aussi l'algorithme SVM.

Nous concluons dans le chapitre 8 en résumant nos principales contributions. Nous y dégageons également quelques pistes de recherche intéressantes autour de nos travaux.

Chapitre 2

Apprentissage automatique et classification supervisée en imagerie satellitaire

Dans plusieurs secteurs d'activité, l'augmentation permanente de la quantité des données rend plus complexes les tâches d'analyse des données par les experts. Avec le développement d'ordinateurs de plus en plus performants, il est possible d'automatiser de tels processus, et d'aider les experts à comprendre de grands ensembles de données très complexes afin de prendre des décisions : c'est l'objet de l'apprentissage automatique. Né des efforts communs des chercheurs issus des domaines de l'informatique et de la statistique, l'apprentissage automatique a pour but d'extraire automatiquement des informations à partir de données, via des approches statistiques ou structurelles. Tandis que les méthodes statistiques, dérivées de la théorie de Vapnik-Chervonenkis [Vapnik, 1996], tentent d'expliquer l'apprentissage d'un point de vue statistique, les méthodes structurelles font usage de l'inférence grammaticale qui dégagent les règles d'un langage formel (sa grammaire) à partir d'exemples positifs ou négatifs, ou de la constitution de prototypes. Elles s'appuient sur la comparaison des représentations structurées avec des représentations prototypes conservées en mémoire : comparaison de graphes, de chaînes et d'arbres. Dans nos travaux, nous nous intéressons particulièrement à l'apprentissage automatique statistique qui regroupe les méthodes d'inférence statistique s'appuyant sur une description des données et/ou de la tâche par des modèles probabilistes. Dans ce contexte, nous présentons dans ce chapitre, la problématique de l'apprentissage automatique, et nous intéressons à son utilisation pour la classification supervisée des images satellitaires en particulier.

2.1 Notion d'apprentissage automatique

A la croisée de plusieurs disciplines telles que les statistiques et l'intelligence artificielle, l'apprentissage automatique (ou *machine learning*) offre un cadre méthodologique et théorique pour l'inférence à partir de données [Hastie et al., 2001; Alpaydin, 2004; Mitchell, 1997]. En effet, il a pour objectif d'extraire et d'exploiter automatiquement l'information présente dans un jeu de données obtenues à partir d'expériences ou d'observations.

Avant 1985, il n'existe quasiment aucune application commerciale de l'apprentis-

sage automatique. Mais actuellement, il est utilisé dans plusieurs domaines de la science, la finance, le commerce et l'industrie [Hastie et al., 2001; Alpaydin, 2004]. Dans le secteur commercial par exemple, les sociétés de vente au détail analysent leurs données de vente passées pour apprendre le comportement de leurs clients, afin d'améliorer la gestion de la relation clientèle. De même en finance, les banques étudient les anciennes données pour construire des modèles qui seront utilisés dans les applications de crédit et de détection de fraudes. Dans le secteur industriel, les robots par exemple, apprennent à optimiser leur comportement pour accomplir des tâches en utilisant le minimum de ressources. En médecine, l'apprentissage automatique est utilisé pour le diagnostic médical, tandis qu'en bioinformatique, il intervient dans l'analyse du génome, pour l'expression des gènes dans différentes situations biologiques.

On trouve également plusieurs applications de l'apprentissage automatique dans le domaine du traitement du signal, des images et de la vidéo [Camastra & Vinciarelli, 2008; Theodoridis & Koutroumbas, 2003]. Il est par exemple utilisé pour la reconnaissance des formes, qui englobe entre autres, la reconnaissance de la parole où l'on doit apprendre l'association d'un signal acoustique à un mot dans un langage donné. Par ailleurs, plusieurs systèmes de vision par ordinateur, depuis ceux de reconnaissance des visages jusqu'aux systèmes qui classifient automatiquement les images microscopiques de cellules, sont développés à l'aide des techniques d'apprentissage automatique, car les systèmes résultants sont plus précis que les programmes développés manuellement. En télédétection, l'apprentissage automatique est utilisé pour l'analyse et la classification des images de télédétection (optiques, radar, aériennes), pour des applications telles que la gestion urbaine, la production des cartes d'occupation du sol (*land cover*) ou encore l'analyse des séismes et autres catastrophes naturelles.

Par ailleurs, l'apprentissage automatique est étroitement lié à l'intelligence artificielle [Russell & Norvig, 1995], dans la mesure où un système intelligent présent dans un environnement qui évolue doit être capable d'apprendre pour s'adapter aux changements de son environnement. Il joue également un rôle clé dans le domaine de la fouille de données.

2.1.1 Fouille de données

La fouille de données (ou *data mining*) est l'application des méthodes d'apprentissage automatique aux grandes bases de données [Weiss & Indurkhy, 1998; Duda et al., 2000]. En effet, elle consiste à rechercher et extraire l'information utile de gros volumes de données, ce qui fait d'elle l'un des maillons de la chaîne de traitement pour l'extraction des connaissances à partir des données (ECD ou *knowledge discovery in databases*) [Fayyad et al., 1996; Frawley et al., 1992].

L'extraction de connaissances dans les bases de données consiste à analyser un ensemble de données brutes de façon à en extraire des connaissances exploitables. En fonction de ses objectifs, un expert du domaine relatif aux données va les sélectionner et utiliser des outils de fouille de données pour construire des modèles expliquant les données. L'expert peut ensuite sélectionner et exploiter les modèles qui représentent un point de vue satisfaisant. Pour mener à bien son activité, l'analyste met à contribution ses connaissances mais aussi un ensemble d'outils regroupés au sein d'un système d'ECD. L'ECD est un processus complexe qui se déroule suivant une suite d'opérations :

- la préparation des données : sélection des descripteurs, constitution d'une table,
-

discrétisation. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population, va dépendre la mise au point des modèles de prédiction,

- l' extraction de connaissances à l'aide d'algorithmes de calcul : c'est la fouille de données, qui produit des motifs potentiellement intéressants. Les données peuvent être stockées dans des entrepôts (*data warehouse*), dans des bases de données distribuées, ou sur Internet (*web mining*). La fouille de données ne se limite pas au traitement des données structurées sous forme de tables numériques, elle offre des moyens pour aborder les corpus en langage naturel (*text mining*), les images (*image mining*), le son (*sound mining*) ou la video et dans ce cas, on parle alors plus généralement de *multimedia mining*.
- et la visualisation et l'interprétation des résultats lors d'interactions avec l'expert.

L'ECD peut être ainsi vue comme le processus alimentant un système à base de connaissances : les connaissances extraites sont stockées dans la base pour être réutilisées dans d'autres applications et mises à jour le cas échéant.

2.1.2 Catégories d'apprentissage automatique

L'apprentissage automatique regroupe un ensemble de méthodes et théories d'analyse des données. Ces méthodes peuvent être classées en fonction du type d'apprentissage utilisé. En effet, les problèmes d'apprentissage peuvent être soit supervisés, soit non supervisés.

Dans l'apprentissage supervisé, le but est de prédire la valeur d'une mesure de sortie, en se basant sur un certain nombre de mesures d'entrée. Le terme "supervisé" est justifié par la présence de la variable de sortie qui guide le processus d'apprentissage (un expert étiquette correctement les valeurs d'entrée). L'on dispose donc d'un échantillon d'exemples constitués de paires entrées/sorties, à partir duquel on veut concevoir une règle générale ou un modèle qui représente la relation entrée/sortie sous-jacente. La valeur de sortie peut être soit un code de classe (sortie qualitative), et on parle alors d'un problème de classification ou encore de discrimination ou d'apprentissage de concepts à partir d'exemples, soit une valeur numérique (sortie quantitative), et dans ce cas, il s'agit d'un problème de régression.

Dans l'apprentissage non supervisé, l'on ne dispose que des données d'entrée et il n'y a aucune mesure de la sortie (il n'y a pas d'expert pour étiquetter correctement les exemples). Le but est donc de décrire comment les données sont organisées ou regroupées, c'est-à-dire, les associations et régularités dans les mesures d'entrée. En statistiques, on parle d'estimation de densité, dont une méthode est le clustering.

Plusieurs algorithmes sont utilisés pour l'extraction automatique d'un modèle à partir d'un échantillon d'observations décrites par des variables. On distingue entre autres, les réseaux de neurones pour un apprentissage supervisé ou non, la méthode des *k*-plus proches voisins pour un apprentissage supervisé, ainsi que les arbres de décision [Breiman et al., 1984; Quinlan, 1993], les méthodes bayésiennes comme le classificateur bayésien naïf, et plus récemment, les machines à vecteurs de support ou SVM (*Support Vector Machines*) [Cortes & Vapnik, 1995], qui permettent le traitement de problèmes de très grande dimension, en particulier dans les cas où les objets sont représentés par un très grand nombre de variables d'entrée.

La tâche la plus étudiée en apprentissage automatique consiste à inférer une fonc-

tion classant des exemples représentés comme vecteurs de traits distinctifs dans une catégorie parmi un ensemble fini de catégories données. Dans la section suivante, nous nous intéressons plus en détail au principe de l'apprentissage automatique supervisé pour la classification.

2.2 Application à la classification supervisée

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets, à partir de certains traits descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée.

2.2.1 Principe

Le principe de l'apprentissage automatique peut être appliqué à la classification. L'information fournie en entrée d'un algorithme d'apprentissage supervisé pour la classification est un échantillon d'objets ou d'exemples décrits par des variables d'entrée et par une variable de sortie. Pour chaque objet, la valeur de la variable de sortie (la classe) est connue et fournie à la méthode. Dans le cas particulier de la classification d'images, les exemples sont des images étiquetées. Le but de l'apprentissage automatique est alors de modéliser les relations entre les entrées et sorties de l'échantillon d'objets. Un tel échantillon est appelé "ensemble d'apprentissage", et sa constitution est réalisée par un expert du domaine, son expérience étant ainsi exploitée implicitement. Le modèle construit est ensuite utilisé pour prendre une décision quant à la classe d'un nouvel objet.

Le problème de la classification peut être décrit par le formalisme suivant : on dispose d'un ensemble $X \subset \mathcal{R}$, constitué de N objets \mathbf{x}_n . Chaque donnée \mathbf{x}_n est caractérisée par D attributs et par sa classe $y_n \in \{C_k, k = 1 \dots K\}$. Le problème consiste alors, en s'appuyant sur l'ensemble des données :

$$X = \{(\mathbf{x}_n, y_n), n = 1 \dots N\} \quad (2.1)$$

à prédire la classe de toute nouvelle donnée $\mathbf{x} \in \mathcal{R}$.

Les algorithmes d'apprentissage utilisent donc les données pour construire un modèle de manière automatique : c'est la phase d'apprentissage. Une fois le modèle construit, on peut vérifier s'il permet une généralisation au-delà de l'ensemble d'apprentissage. Pour cela, on utilise un échantillon de test, contenant des objets distincts de ceux de l'ensemble d'apprentissage. La précision du modèle est généralement exprimée par son taux d'erreur ou son taux de bonne reconnaissance sur cet échantillon de test. C'est la phase de prédiction, qui consiste à utiliser le modèle construit pour attribuer une classe à un nouvel objet $\mathbf{x} \in \mathcal{R}$.

Pour une bonne généralisation, il est nécessaire que la complexité du modèle et celle de la fonction générant les données correspondent. Si le modèle construit est trop simple, on parle de sous-apprentissage et on exprime la composante dominante de l'erreur par le biais. D'autre part, un apprentissage trop précis peut ne pas être généralisable. Ce phénomène est appelé sur-apprentissage, et l'erreur est exprimée sous la forme d'une variance. Les deux sources d'erreur : le biais et la variance, sont liées à la complexité du modèle, mais de manière opposée. On parle alors du compromis biais / variance

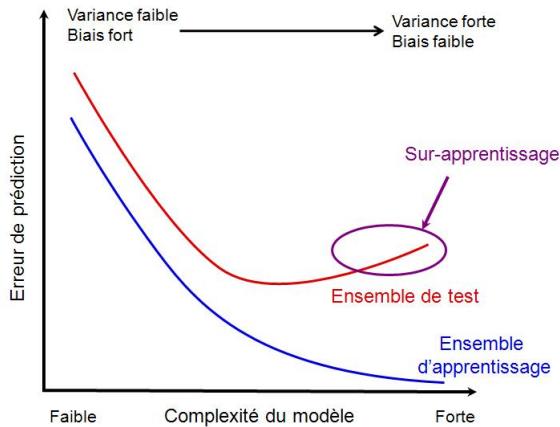


FIG. 2.1 – Comportement de l'erreur de prédiction lorsque la complexité du modèle varie.

[Hastie et al., 2001]. La figure 2.1 montre le comportement de l'erreur de prédiction lorsque la complexité du modèle varie.

2.2.2 Etat de l'art en imagerie satellitaire

Depuis plusieurs années, les capteurs optiques fournissent aux thématiciens de la télédétection pour l'observation de la Terre, des images de plus en plus nombreuses et de plus en plus riches. Pour exploiter correctement et efficacement la quantité d'informations contenue dans les images satellitaires, et en particulier pour produire des cartes thématiques de haute qualité, des méthodes de traitement et d'analyse automatiques des images ont été développées. Dans la littérature, des efforts importants ont été consacrés aux techniques d'apprentissage automatique pour la classification. Et actuellement, ces méthodes font encore l'objet de nombreux travaux. En effet, la classification est vue comme un processus fondamental en télédétection, se situant au coeur de la chaîne de transformation de l'image satellitaire en un produit géographique utilisable : elle fournit une couche classifiée, à mettre en correspondance avec des informations d'autres sources dans un Système d'Information Géographique (SIG), ou bien à utiliser comme donnée de base dans la construction d'un modèle dynamique de l'environnement.

Les méthodes de classification supervisée utilisent en général deux types d'approches : celles qui sont basées sur le pixel (*pixel-based approach*) et celles qui sont basées sur la région (*region-based approach*), appelée objet dans certains travaux (*object-based approach*). Une image étant un tableau de pixels, les techniques classiques utilisent l'approche basée sur le pixel [Jensen, 1986; Casals-Carrasco et al., 2000]. Dans les méthodes supervisées, chaque pixel de l'image est attribué à une classe définie au préalable, principalement à partir de ses caractéristiques de bas-niveau. Les méthodes utilisant des fenêtres de pixels [Keramitsoglou et al., 2006] sont souvent associées à ce type d'approches. Quant à l'approche par régions, elle considère l'image comme un ensemble de régions significatives (ou objets) obtenues via un partitionnement de l'image effectué au préalable (une segmentation par exemple), et classifie ainsi des régions homogènes de l'image, plutôt que chaque pixel séparément [Baraldi & Parmiggiani, 1990; Kusaka et al., 1990]. Par rapport à la première approche, celle-ci présente donc l'avantage de tenir compte de l'arrange-

ment spatial, c'est-à-dire de l'aspect contextuel, qui peut être exploité par la suite pour améliorer les performances de classification. En imagerie satellitaire, suivant la résolution des images et l'application souhaitée, l'une ou l'autre approche peut être appropriée.

La diversité des méthodes de classification présentes dans la littérature découle principalement de la variété des algorithmes d'apprentissage automatique et des types d'information utilisés. En effet, Wilkinson [Wilkinson, 2005] montre que ces deux éléments de la procédure de classification sont parmi ceux qui ont le plus évolué ces dernières années.

2.2.2.1 Algorithmes d'apprentissage

La procédure statistique de classification basée sur le Maximum de Vraisemblance (MV) et exploitant principalement les propriétés spectrales, est la plus utilisée dans les approches basées sur le pixel. Elle est très rapidement devenue une méthode standard, plusieurs dérivés de cet algorithme développés par la suite ayant pour but de dépasser les performances du classificateur MV. Par exemple, dans [Gorte & Stein, 1998], les auteurs proposent une méthode de classification bayésienne au niveau du pixel, basée sur une extension itérative de la méthode du Maximum a Posteriori (MAP), qui elle-même, est une généralisation de celle du maximum de vraisemblance lorsque la distribution a priori n'est pas uniforme. A chaque itération, les probabilités a posteriori pour chaque classe sont utilisées non seulement pour la classification, mais aussi pour obtenir les estimations des surfaces de chaque classe. Ces dernières sont ensuite utilisées pour calculer un ensemble de probabilités a priori actualisé qui sera utilisé dans la prochaine itération. La méthode, illustrée sur une image Landsat TM, permet de mieux identifier des classes telles que les zones industrielles, l'agriculture, les zones résidentielles, et la végétation naturelle, par rapport à la méthode classique du maximum de vraisemblance.

De manière générale, dans la plupart des travaux de classification supervisé, et même dans les approches par régions, la classification basée sur le maximum de vraisemblance est comparée aux méthodes proposées, afin de juger des performances de ces dernières [Casals-Carrasco et al., 2000; Dean & Smith, 2003; Yan et al., 2006]. Gao et ses collègues [Yan et al., 2006] par exemple, comparent une classification supervisée basée sur le maximum de vraisemblance et opérée sur chaque pixel, et une classification orientée objet. Cette dernière partitionne d'abord l'image à l'aide d'une segmentation multi-résolution par croissance de régions, puis utilise un classificateur soft basé sur les plus proches voisins. L'évaluation de ces deux méthodes utilisées pour cartographier l'occupation des terres dans une région de feu de charbon en Chine, montre que la classification orientée objet présente un taux global de bonne classification supérieur de 37% à celui de l'autre classification. De même, Frizelle et Moody [Frizelle & Moody, 2001] opposent un classificateur basé sur le maximum de vraisemblance et un classificateur basé sur les réseaux de neurones, en les utilisant pour caractériser l'occupation des terres comme des champs continus représentant soit les proportions des classes, soit la certitude de classification.

Par ailleurs, de nombreux autres travaux de classification présents dans la littérature sollicitent les réseaux de neurones [Bischof et al., 1992; Serpico & Roli, 1995; Peng et al., 2003]. Dans [Bischof et al., 1992], les auteurs présentent une application de réseaux à trois couches avec rétro-propagation, pour la classification d'images Landsat TM, basée sur une approche pixel. Ils montrent que cette méthode donne de meilleures performances que celles obtenues avec un classificateur gaussien au sens du maximum de vraisemblance. Certaines études telles que [Peng et al., 2003], étendent l'utilisation des réseaux

de neurones pour la classification en les combinant avec d'autres types de traitements, afin d'en améliorer l'efficacité. En effet, Peng, Li et Su proposent une méthode de classification des images de télédétection basée sur la théorie de l'évidence de Dempster-Shafer [Shafer, 1976] et les réseaux de neurones. L'image d'origine est d'abord lissée, puis un réseau de neurones avec rétro-propagation est utilisé pour apprendre et classifier séparément l'image originale et l'image lissée. Le résultat de classification final est obtenu en fusionnant à l'aide de la théorie de l'évidence, les deux classifications produites par le réseau de neurones. Les auteurs soutiennent que cette méthode améliore de manière notable, les performances de classification données par l'utilisation des réseaux de neurones uniquement.

Les arbres de décision [Hansen et al., 1996], encore appelés arbres de classification, sont une alternative intéressante aux approches traditionnelles de classification de l'occupation des sols. En effet, les arbres fournissent une méthode de classification hiérarchique et non-linéaire, et sont bien adaptés pour manipuler des données d'apprentissage non-paramétriques, aussi bien que des données catégoriques ou manquantes. Hansen et ses collègues [Hansen et al., 1996] montrent que les performances obtenues avec les arbres de décision sont comparables à celles données par le maximum de vraisemblance. Par ailleurs, ils affirment que les arbres peuvent être utilisés pour réduire la dimensionnalité des données et trouver des métriques utiles pour faire la distinction entre les différents types d'occupation du sol. Pal et Mather [Pal & Mather, 2003] quant à eux, utilisent des images de deux zones géographiques différentes, obtenues avec les capteurs Multispectral Landsat ETM+ et Hyperspectral DAIS, pour juger des performances des arbres de décision à une et plusieurs variables, pour la classification de l'occupation des sols. Le degré d'exactitude de la classification réalisée par les arbres de décision est comparé aux résultats obtenus avec les classificateurs basés sur les réseaux de neurones artificiels avec rétro-propagation et le maximum de vraisemblance. Comme dans [Hansen et al., 1996], les performances du classificateur par arbres de décision sont comparables à celles des deux autres classificateurs, sauf pour les données de grande dimension pour lesquelles les arbres de décisions sont moins efficaces que les réseaux de neurones et le maximum de vraisemblance.

Pour traiter des données de grande dimension, les SVM (Machines à vecteurs de support ou *Support Vector Machines*) constituent une méthode de classification supervisée particulièrement bien adaptée. Les résultats produits par cette méthode sont souvent comparés à ceux d'autres algorithmes tels que les k-plus proches voisins, les réseaux de neurones artificiels ou encore les arbres de décision univariés [Melgani & Bruzzone, 2004; Pal & Mather, 2004; Zammit et al., 2007]. Et, comme Melgani et Bruzzone l'ont conclu, les SVM s'avèrent être une alternative valide et efficace aux approches conventionnelles de reconnaissance des formes pour la classification de données de télédétection hyperspectrales par exemple.

D'autres méthodes de classification, moins utilisées dans la littérature, sont basées sur les fonctions densité de probabilité n -dimensionnelles [Cetin et al., 1993], ou encore l'analyse de forme spectrale [Carlotto, 1998] pour les données multispectrales, la forme spectrale d'une classe étant un vecteur d'attributs binaires qui décrivent les valeurs relatives entre les bandes spectrales.

2.2.2.2 Types d'information

La variété des méthodes de classification provient aussi de la diversité des informations utilisées. Il s'agit des différents types d'information extraites des images, de l'utilisation de données multisources ou encore de l'information exogène. Nous nous intéressons dans cette analyse uniquement aux données provenant des images.

Les primitives triviales utilisées pour la classification sont spectrales ou basées sur les niveaux de gris. Dans le cas d'images à plusieurs bandes spectrales, ces dernières peuvent être combinées pour mettre en évidence des thèmes particuliers de l'image : ce sont les néocanaux. Les plus populaires sont les indices de végétation, qui sont utilisés pour déterminer le taux de végétation dans l'image. Dans leurs travaux de classification d'images satellitaires, [Kusaka et al., 1990] et [Unsalan & Boyer, 2004] utilisent l'indice de végétation normalisé (NDVI ou *Normalized Difference Vegetation Index*) comme une information multispectrale additionnelle.

D'autres primitives ont été extraites des images, et sont généralement combinées aux caractéristiques spectrales afin d'améliorer les performances de classification. Parmi elles, les mesures de texture sont les plus utilisées [Song et al., 2006; He & Collet, 1999; Gong et al., 1992]. Dans [He & Collet, 1999] par exemple, les auteurs proposent une procédure de classification d'images multispectrales de type SPOT XS, qui introduit des caractéristiques texturelles issues de matrices de co-occurrence des niveaux de gris, dans une classification conventionnelle basée sur le pixel, et implémentée à l'aide des réseaux de neurones. De même, Aksoy et ses collègues [Aksoy et al., 2005] fusionnent les caractéristiques spectrales, texturelles basées sur les filtres de Gabor, et des données ancillaires, pour classifier des scènes Landsat, dans un cadre bayésien. Song, Li et Yang [Song et al., 2006] quant à eux, proposent une mesure de texture appelée *Local Binary Pattern*, dont la version à plusieurs variables est utilisée dans le processus de classification d'images de télédétection multispectrales. Les auteurs montrent que cette méthode peut améliorer les performances de classification de manière significative, par rapport à la classification spectrale.

D'autre part, les caractéristiques spatiales et structurelles, qui tiennent compte de l'information contextuelle, peuvent également être utilisées pour la classification, généralement avec des approches basées sur la région. Dans [Kusaka et al., 1990] par exemple, les auteurs classifient une image SPOT multispectrale en utilisant les caractéristiques spectrales et spatiales de ses régions primitives obtenues via une technique de segmentation basée sur les contours. Ils montrent que l'ajout des caractéristiques spatiales telles que le facteur de forme, la compacité ou encore la variation de luminosité des régions primitives, améliore la qualité de la classification. Gong et Howarth [Gong & Howarth, 1990] décrivent une méthodologie pour incorporer l'information structurelle dans les procédures de classification conventionnelles. Cette information est contenue dans une image de densité des contours, obtenue en appliquant un filtre passe-haut Laplacien et en extrayant les contours. En utilisant l'image de densité des contours comme une bande additionnelle dans le classificateur de Mahalanobis, cette méthode améliore de 9.5%, les performances de la classification conventionnelle par maximum de vraisemblance. De même, Dell'Acqua et ses collègues [Dell'Acqua et al., 2004] combinent l'information spectrale et l'information spatiale obtenue à l'aide des profils morphologiques basés sur la morphologie mathématique, pour classifier les images hyperspectrales à haute résolution.

Certaines méthodes de classification font usage de plusieurs images issues de cap-

teurs différents, ou du même capteur mais à des résolutions différentes. C'est dans ce contexte que s'effectuent les tâches de fusion de données optiques et radar pour l'analyse d'images. Cependant ici, nous ne nous intéressons qu'aux images issues de capteurs optiques uniquement. Unsalan et Boyer [Unsalan & Boyer, 2004] par exemple, proposent une méthode de classification du développement des sols à partir d'images à haute résolution, basée sur l'utilisation de primitives spectrales et structurales hybrides. L'information structurale est extraite d'une image Ikonos Panchromatique à 1 m de résolution, sous la forme de lignes droites photométriques et leur arrangement spatial dans un voisinage, tandis que la réponse multispectrale est issue d'une image Ikonos à 4 m de résolution, et est exprimée par le NDVI et sa forme linéarisée. Les caractéristiques hybrides, obtenues en combinant l'information structurale et la réponse multispectrale, permettent de mieux distinguer les régions urbaines des régions rurales que chaque caractère séparément. Dans le même ordre d'idées, le système GeoIRIS [Shyu et al., 2007], développé pour la recherche et l'indexation par le contenu, utilise des informations provenant de bases de données hétérogènes. En effet, dans leur système, les auteurs extraient des caractéristiques d'une image panchromatique à haute résolution (0.6 - 1 m) et d'une image satellitaire multispectrale (2.4 - 4 m), qui ont été fusionnées pour créer une image multispectrale à 0.6 - 1 m de résolution, avec la même résolution spatiale pour tous les canaux, y compris l'infrarouge.

D'autres facteurs liés à la nature même du système de classification, peuvent être pris en compte pour traduire la multiplicité des méthodes de classification dans la littérature. En effet, il existe par exemple des systèmes de classification floue [Zhang & Foody, 1998] qui "arrondissent" les résultats donnés par les classificateurs durs, et les systèmes multi-classificateurs [Wilkinson et al., 1995] qui intègrent les sorties de plusieurs classificateurs de base.

2.3 Conclusions

La reconnaissance de la couverture du sol à partir des classifications automatiques est l'une des recherches méthodologiques importantes en télédétection. Ces méthodes de classification utilisent des algorithmes d'apprentissage pour construire de manière automatique un modèle de classification à partir d'images étiquetées. L'information contenue dans ce modèle permet de prendre une décision quant à la classe d'une nouvelle image. De cette étude, nous retenons donc que la diversité des méthodes de classification dépend principalement des algorithmes d'apprentissage utilisés. Cependant, elle est également fonction des choix possibles pour les différents éléments du processus complet de classification, entre autres le choix des primitives. Dans le chapitre suivant, nous décrivons en détail chacune des étapes de la chaîne de classification supervisée.

Nous notons en outre, les fortes occurrences de la méthode classique de classification basée sur le maximum de vraisemblance dans les travaux existants, méthode très souvent utilisée pour effectuer un comparatif avec les approches proposées, et ainsi juger des performances de ces dernières.

Chapitre 3

Processus de classification automatique supervisée

Dans nos travaux, nous nous intéressons à un problème actuel lié à la classification automatique d'images satellitaires à l'aide de classes définies au préalable, pour les applications de type cartographie de l'occupation du sol. Nous présentons donc dans ce chapitre, une description des différentes étapes de la chaîne de classification supervisée, depuis l'extraction des primitives jusqu'aux méthodes d'optimisation de la classification. Nous exposons tout d'abord, les traitements de bas-niveau opérés sur l'image, car les performances de la classification dépendent entre autres, de cette information portée par des attributs numériques directement calculés sur l'image. Nous nous servirons ensuite de cette information comme d'une base pour décrire le contenu sémantique de l'image.

3.1 Du pixel à la région : la segmentation

Les images peuvent être analysées au niveau de la scène, suivant une approche basée sur le pixel, ou encore une approche basée sur la région. Cette dernière approche consiste à extraire des caractéristiques des régions de l'image, obtenues via une segmentation.

La segmentation a pour but de partitionner l'image en régions homogènes, grâce à un critère prédéfini. Idéalement, cette homogénéité est sémantique, de telle sorte que les différentes régions de la segmentation correspondent aux différents objets de l'image. Cependant, en pratique, les algorithmes de segmentation produisent des régions homogènes suivant des attributs de bas-niveau.

Il existe des centaines d'algorithmes de segmentation dans la littérature. Pal et Pal [Pal & Pal, 1993] proposent une critique de quelques algorithmes de segmentation. Certaines méthodes, basées sur les histogrammes, sont de mise en oeuvre assez simple et ont des performances souvent réduites, car elles ne tirent pas profit de l'aspect spatial de l'information d'image. L'idée générale de ces méthodes consiste à isoler les pics de l'histogramme : à une dimension, on procède donc à des seuillages [Otsu, 1979] ou des multi-seuillages [Nakagawa & Rosenfeld, 1979; Taxt et al., 1989], tandis qu'à n dimensions, on opère des classifications [Dubuisson, 1990].

Certaines méthodes de segmentation, telles que la croissance de régions (*region growing*) [Adams & Bishof, 1994] ou le partage et la réunion de régions (*split and merge*), opèrent directement sur les pixels de l'image. Par ailleurs, d'autres techniques basées sur la détection de contours, qui essaye de localiser les points de changement brusque

d'intensité dans l'image, cherchent à exploiter le fait qu'il existe une transition détectable entre deux régions connexes. Ces deux types d'approches de la segmentation sont particulièrement détaillées dans [Bolon et al., 1995]. Mueller et ses collègues [Mueller et al., 2004], quant à eux, proposent une méthode de segmentation basée à la fois sur les régions et les contours, pour l'extraction de champs d'agriculture dans les images satellitaires à haute résolution.

Plusieurs travaux sur la segmentation des images utilisent les modèles d'interaction spatiale tels que les champs de Markov (*Markov Random Fields* ou MRF) ou de Gibbs (*Gibbs Random Fields* ou GRF), pour modéliser les images [Derin & Elliot, 1987; Jain, 1981]. Plus récemment, Deng et Clausi [Deng & Clausi, 2004], présentent une méthode classique de segmentation basée sur les MRF, dans laquelle le poids du terme de régularisation varie selon les itérations de l'algorithme. De même, dans [Poggi et al., 2005], les auteurs utilisent les champs de Markov structurés en arbre, pour améliorer les performances de la segmentation. Par ailleurs, les réseaux de neurones, qui permettent d'avoir la sortie en temps réel grâce à leur aptitude à effectuer des traitements en parallèle, ont montré de bonnes performances pour la segmentation d'images, même quand le niveau de bruit est très élevé. Blanz et Gish [Blanz & Gish, 1990] par exemple, ont utilisé un réseau à trois couches pour la segmentation des images, dans lequel le nombre de neurones dans la couche d'entrée dépend du nombre de primitives d'entrée pour chaque pixel, et le nombre de neurones dans la couche de sortie est égal au nombre de classes.

On distingue en outre, les méthodes basées sur le partitionnement de graphes, comme les Normalized cuts [Shi & Malik, 2000]. L'image est modélisée par un graphe pondéré non dirigé, chaque pixel étant un noeud du graphe et un arc étant formé entre toute paire de pixels. Les approches théoriques d'ensembles flous introduits par Zadeh [Zadeh, 1965], mais aussi les méthodes de segmentation multi-échelle [Salembier & Garrido, 2000; Vanhamel et al., 2003] sont d'autres techniques de segmentation non sans intérêt.

De toutes ces méthodes, aucune d'elles ne peut être considérée comme performante sur toutes les images, tout comme toutes les méthodes ne peuvent avoir la même performance sur un type d'image particulier. Ci-dessous, nous présentons des méthodes de segmentation souvent utilisées pour l'analyse et l'interprétation des images de télédétection. Parmi les algorithmes de segmentation les plus utilisés dans la littérature pour l'annotation et la recherche d'images par le contenu, on trouve la Ligne de Partage des Eaux (LPE ou *watershed*) et les *Normalized cuts*.

Watershed : Le watershed [Beucher & Meyer, 1993] utilise la description des images en termes géographiques. En effet, cet algorithme considère qu'une image peut être perçue comme un relief si l'on associe le niveau de gris de chaque point à une altitude. Il est alors possible de définir la ligne de partage des eaux comme étant la crête formant la limite entre deux bassins versants. Appliqué à une image de gradient, cet algorithme fournit une partition de l'image où une région est associée à chaque minimum du gradient. La prolifération de minima dans l'image de gradient conduit systématiquement à une sur-segmentation de l'image originale. La figure 3.1 présente une image multispectrale Pelican à 50 cm de résolution (figure 3.1(a)) et sa segmentation obtenue à l'aide du watershed (figure 3.1(b)). On remarque la présence de plusieurs petites régions au niveau des frontières et au niveau des lignes sur la route. De même, les zones de végétation ont tendance à être découpées en plusieurs régions homogènes. Une amélioration de la segmentation est possible en appliquant une fermeture morphologique à l'image de

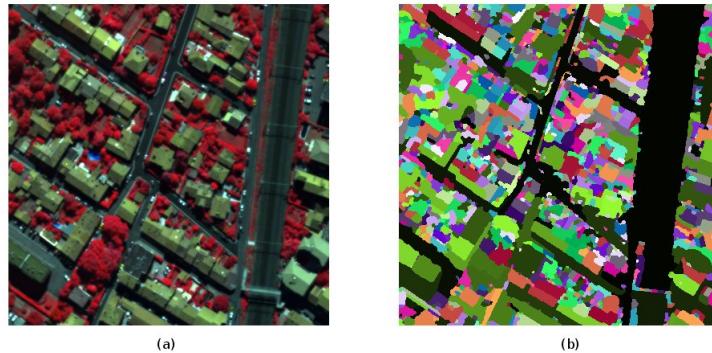


FIG. 3.1 – Test de segmentation par l’algorithme Watershed. (a) Image Pelican à 50 cm de Toulouse. (b) Segmentation obtenue : 1300 régions.

gradient, pour réduire le nombre de minima. La taille de l’élément structurant utilisé influence directement le nombre final de régions obtenues.

Partitionnement de graphes : L’algorithme des *Normalized cuts* [Shi & Malik, 2000] est courant dans la littérature pour l’annotation et la recherche d’images. Proposée par Shi et Malik, cette approche traite la segmentation comme un problème de partitionnement de graphe. Le critère de segmentation du graphe, “normalized cut”, mesure la dissimilarité totale inter-groupes, ainsi que la similarité totale intra-groupe. Cette méthode de segmentation produit quelquefois des régions de petite taille et instables. Dans [Barnard et al., 2003] et [Duygulu et al., 2002], les régions plus petites qu’un certain seuil ne sont pas utilisées pour l’annotation. Une technique de calcul efficace basée sur un système généralisé de valeurs propres peut être utilisée pour optimiser ce critère. Dans la même catégorie d’approche, on trouve les hypergraphes. En effet, l’image peut aussi être représentée par un hypergraphe [Berge, 1973; Bretto & Ubéda, 1996], qui est une généralisation du graphe dans la mesure où les relations entre les pixels n’y sont pas exclusivement binaires.

Avec un concept légèrement similaire aux *Normalized cuts*, la méthode de segmentation proposée par Lersch *et al.* [Lersch et al., 1996] et utilisée dans [Watanabe et al., 2002], est basée sur le calcul de l’arbre recouvrant de poids minimal du graphe de pixels de l’image. L’algorithme de coupure de l’arbre est basé sur un moyennage spectral, en commençant à chaque itération, par l’arête de poids minimum. Ainsi, les arcs de poids supérieur à un certain seuil sont éliminés, et les arcs restants ont tendance à connecter les pixels de couleur similaire, constituant ainsi des régions potentielles de l’image. La figure 3.2(a) montre un exemple de segmentation de l’image de la figure 3.1(a) par cet algorithme. Visuellement, cette segmentation est plus intéressante que le watershed car les régions obtenues correspondent mieux aux objets de l’image (surtout végétation et bâtiments). Cependant le nombre de régions est élevé (1800 dans ce cas), le seuil ayant été fixé ici à 45. Plus la valeur du seuil augmente, plus la segmentation est grossière et donc, le nombre de régions diminue.

Mean shift : Une autre méthode de segmentation est Le Mean Shift [Comaniciu & Meer, 2002], qui est une procédure itérative de montée de gradient, utilisée pour estimer les modes d’une fonction densité associée à un échantillon d’observations. Soit une variable

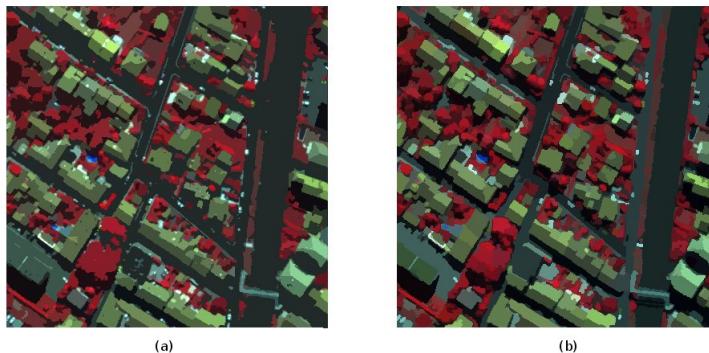


FIG. 3.2 – Autres tests de segmentation sur l'image de la figure 3.1(a). (a) Image segmentée par la méthode de Lersch basée sur le calcul de l'arbre recouvrant de poids minimal : 1800 régions. Le seuil a été fixé à 45. (b) Image segmentée par Mean shift : 1393 régions. La taille de la région minimale est de 20 pixels.

aléatoire x avec la distribution $p(x)$, l'idée est d'estimer directement le gradient de $p(x)$ et d'en chercher les zéros, sans estimer au préalable $\hat{p}(x)$. L'optimisation de montée de gradient consiste en l'itération du Mean Shift jusqu'à la convergence. En contrôlant la taille de la région minimale, il est possible de contrôler le nombre de régions de l'image segmentée. Un exemple de segmentation par mean shift est donné dans la figure 3.1(b), où il ya moins de petites régions car la taille de la région minimale est de 20 pixels.

D'autres méthodes de segmentation sont aussi utilisées. Par exemple, [Aksoy et al., 2005] applique un algorithme de type "décomposition et fusion" de régions (*split and merge*), basé sur des opérateurs morphologiques, pour segmenter une image de pixels étiquetés.

3.2 Extraction de primitives

Pour toute opération d'analyse et de fouille d'images, le type de primitives utilisées, encore appelés attributs ou caractéristiques, est essentiel, car le succès des opérations ultérieures dépend de cette information de bas-niveau, extraite de l'image. Dans la littérature, ces attributs caractérisent des pixels, mais aussi des régions ou des objets de l'image, et peuvent être classés en trois catégories : spectrales, texturelles et géométriques.

3.2.1 Primitives spectrales

La littérature propose outre le niveau de gris de chaque pixel pour chaque bande spectrale (vecteur multispectral), les caractéristiques statistiques et ce que l'on appelle les néocanaux, résultant de traitements élaborés à partir de plusieurs bandes spectrales.

Les caractéristiques statistiques sont en général la moyenne et l'écart-type (moments statistiques du premier et second ordre), calculés dans un certain voisinage. Les moments statistiques d'ordre supérieur tels que l'asymétrie (skewness) et l'aplatissement (kurtosis) peuvent aussi être utilisés pour décrire l'image. Soit I , une image en niveaux de gris, sa

moyenne s'exprime de la manière suivante :

$$m_1 = \frac{1}{N_l N_c} \sum_{l=1}^{N_l} \sum_{c=1}^{N_c} I_{lc} \quad (3.1)$$

et son écart-type, qui mesure la dispersion des échantillons autour de la moyenne, est donné par :

$$m_2 = \sqrt{\frac{1}{N_l N_c} \sum_{l=1}^{N_l} \sum_{c=1}^{N_c} (I_{lc} - m_1)^2} \quad (3.2)$$

Cependant, il y a un risque de perte d'information lorsque le voisinage est grand. Une alternative est alors d'utiliser les histogrammes [Swain & Ballard, 1991].

Les néocanaux sont des analyses multivariées, très utilisées dans le traitement des images satellitaires. Il s'agit souvent d'opérations mathématiques visant soit à réduire la somme d'informations (en codage RVB par exemple, on ne peut visualiser que 3 canaux en même temps), soit la mise en évidence de thèmes particuliers (végétation, sols, ...). Les indices de végétation, par exemple, sont utilisés pour déterminer le taux de végétation dans l'image (*Leaf Area Index* ou LAI) et ceci, pixel par pixel. Il en existe plusieurs, séparables en 3 catégories : les indices intrinsèques (NDVI), les indices liés aux variations du sol (PVI, WDVI) et les indices liés aux propriétés de l'atmosphère tels que ARVI et GEMI [Rondeaux et al., 1996]. Ils diffèrent par leur capacité à bien estimer le LAI, et par leur sensibilité à la brillance des sols ou aux effets de l'atmosphère [Bannari et al., 1997]. Le plus utilisé est le NDVI (*Normalized Difference Vegetation Index*), défini par :

$$NDVI = \frac{PIR - R}{PIR + R} \quad (3.3)$$

où PIR et R sont respectivement les réflectances (mesure de la puissance réfléchie sur la puissance reçue) dans les canaux proche infrarouge et rouge. Il a été élaboré à partir des canaux 5 et 7 de Landsat MSS, puis étendu à d'autres types de capteurs. La réponse spectrale d'un couvert végétal dense est forte dans les longueurs d'onde proche infrarouge et faible dans les longueurs d'onde rouge (la chlorophylle absorbe le rayonnement incident de la partie visible du spectre électromagnétique), alors que la réponse spectrale d'un couvert clairsemé est inverse. La différence normalisée permet de rendre compte de ces deux phénomènes sur une même image. Le néocanal résultant présente un gradient croissant d'activité végétale allant du noir signifiant l'absence de couverture, au blanc qui rend compte d'une activité chlorophyllienne très élevée. La figure 3.3 montre l'exemple d'une image SPOT à 20 m de résolution et l'image NDVI correspondante.

Par ailleurs, au lieu d'utiliser les caractéristiques statistiques d'une part et le NDVI d'autre part, Ünsalan et Boyer [Ünsalan & Boyer, 2004] ont eu l'idée de les utiliser conjointement. En effet, en plus du NDVI, ils calculent une version linéarisée de cet index notée θ . C'est une mesure d'angle qui évite le problème de saturation du NDVI, présent lorsque le degré de végétation augmente : $\theta = \frac{4}{\pi} \arctan(NDVI)$. Les caractéristiques multispectrales considérées sont alors les moyenne, variance, skewness et kurtosis de ces deux grandeurs.

Il existe d'autres indices, tels que les indices de brillance et de clarté, qui permettent d'identifier les sols nus en les différenciant de la roche, ou encore l'indice de bâti décrit

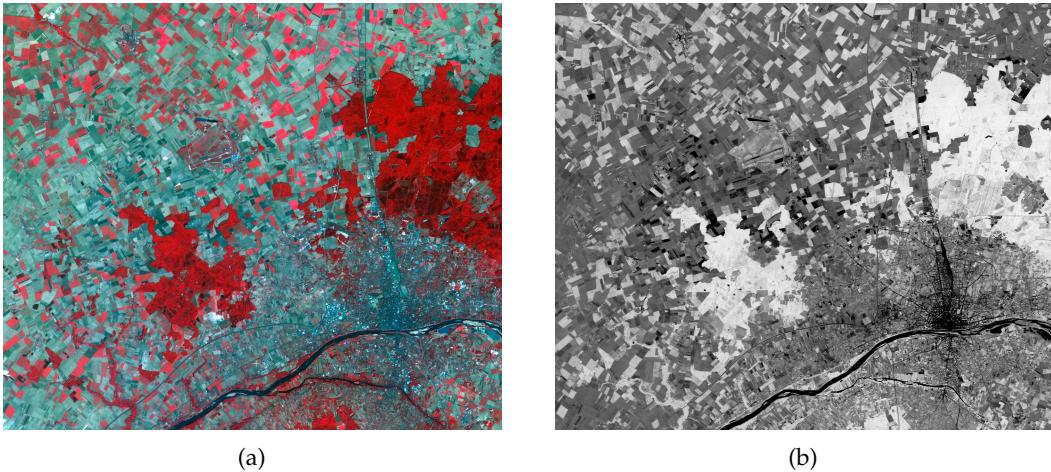


FIG. 3.3 – Exemple d'une image SPOT à 20 m de résolution (figure 3.3(a)) et l'image NDVI correspondante (figure 3.3(b)). Cette dernière présente un gradient croissant d'activité végétale allant du noir signifiant l'absence de couverture, au blanc qui rend compte d'une activité chlorophyllienne très élevée.

dans [Abdellaoui & Rougab, 1997] pour caractériser le bâti. L'indice de brillance IB et l'indice de bâti noté ISU s'expriment par :

$$IB = \sqrt{R^2 + PIR^2} \quad \text{et} \quad ISU = A - B \cdot \frac{R}{PIR} \quad (3.4)$$

où A et B sont des constantes égales respectivement à 100 et 25 pour une image HRV de SPOT.

Dans les approches par régions ou par objets, les moyennes ou médianes de ces valeurs sont utilisées. Par exemple, pour les images SPOT, le niveau de gris moyen de la région dans les bandes 2 et 3 est utilisé pour calculer l'indice de végétation [Kusaka et al., 1990; Baraldi & Parmiggiani, 1990].

3.2.2 Primitives de texture

Selon Forsyth et Ponce [Forsyth & Ponce, 2003], la texture est un phénomène très courant dans les images, facile à reconnaître, mais difficile à définir [Forsyth & Ponce, 2003]. Il en existe d'ailleurs plusieurs définitions différentes dans la littérature, entre autres, plusieurs occurrences d'un élément de base de l'image (souvent appelé *texton*), organisées d'une manière particulière, ou encore une structure périodique détectable avec des outils d'analyse fréquentielle tels que la transformée de Fourier. Dans ses travaux de thèse, Coggins [Coggins, 1982] fournit un récapitulatif des différentes définitions de la texture dans le domaine de la vision par ordinateur.

La texture peut être utilisée pour faire la différence entre deux objets de même radiométrie. Et parce que la texture est liée aux propriétés physiques des objets, il est possible d'identifier, au moins en partie, le contenu d'une région grâce à sa texture. En effet, la texture est caractérisée par des variations dans une image, généralement causées par une variation physique fondamentale dans la scène (comme par exemple, les vagues dans

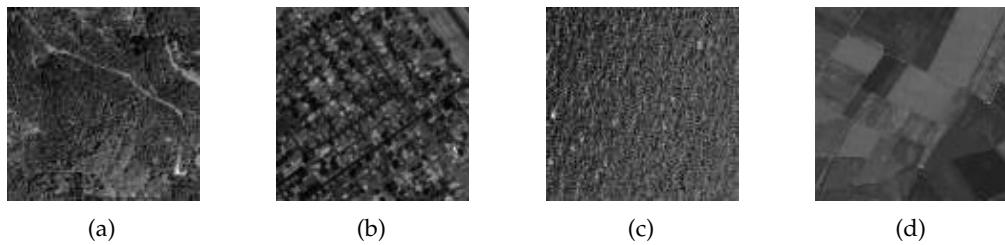


FIG. 3.4 – Exemples d’images texturées, extraites de scènes SPOT5 ©CNES : (a) forêt, (b) ville, (c) mer et (d) champs.

l’eau). La modélisation de cette variation physique étant très difficile, la texture est souvent caractérisée par les variations bidimensionnelles des intensités de l’image. La figure 3.4 montre des exemples de textures dans les images satellitaires.

Dans la littérature, il existe plusieurs comparaisons expérimentales des algorithmes d’extraction de textures [Singh & Singh, 2002; Grigorescu et al., 2000]. En exploitant les propriétés des textures, Tuceryan et Jain [Tuceryan & Jain, 1998] proposent une taxonomie des méthodes d’extraction de textures, et les classifient en géométriques, structurales, statistiques, basées sur les modèles et basées sur le traitement du signal. Par la suite, Randen et Husoy [Randen & Husoy, 1999] concluent que la plupart des travaux utilisent les trois dernières techniques.

La texture implique la distribution spatiale des niveaux de gris. L’utilisation des caractéristiques statistiques est par conséquent l’une des premières méthodes proposées dans la littérature. Parmi elles, on trouve les très populaires matrices de cooccurrence [Haralick et al., 1973], utilisées par exemple dans [Steinnocher et al., 2003] pour la dérivation d’un détecteur de zones urbaines, simple et efficace, et les caractéristiques d’autocorrelation. Par ailleurs, la fonction d’autocorrelation d’une image peut être utilisée pour évaluer la quantité de régularité ainsi que la finesse ou la grossièreté de la texture présente dans l’image. Plus simplement, les textures peuvent aussi être décrites par des histogrammes [Lowitz, 1983].

Plusieurs méthodes d’analyse de texture reposent sur les techniques de traitement du signal. Dans ces approches, l’image texturée est soumise à une transformation linéaire, un filtre ou un banc de filtres. Les méthodes basées sur la transformée de Fourier permettent d’extraire l’orientation locale et la périodicité, mais ne sont pas localisées, contrairement aux approches basées sur les ondelettes [Mallat, 1989], filtres de Gabor [Dunn & Higgins, 1995; Dunn et al., 1994], et filtres miroirs en quadrature (*Quadrature Mirrors Filters* ou QMF) [Mallat, 2003], qui tiennent compte de la localisation dans le domaine spatial. Hsu, Calway et Wilson [Hsu et al., 1993], font usage de l’analyse de Fourier pour la description de textures, ceci en utilisant la transformée de Fourier multirésolution pour la segmentation de l’image et la synthèse de textures, tout en tenant compte de la déformation géométrique typique des images de textures naturelles. Par ailleurs, Myint et ses collègues [Myint et al., 2004] présentent une caractéristique de texture basée sur les ondelettes pour la classification des types d’utilisation des terres des zones urbaines et suburbaines en imagerie satellitaire à haute résolution, et la comparent aux descripteurs basés sur les fractales ou aux matrices de cooccurrences.

Les méthodes d’analyse de texture, basées sur la construction d’un modèle de l’image,

regroupent entre autres les champs de Markov (*Markov Random Fields* ou MRF) et la dimension fractale, très utilisés pour la classification des images de télédétection. Utilisés dans [Cross & Jain, 1983], les champs de Markov sont capables de capturer l'information contextuelle locale (spatiale) dans une image. Ces modèles supposent que l'intensité de chaque pixel de l'image dépend uniquement de l'intensité des pixels de son voisinage. Les attributs issus de la modélisation de l'image par les champs de Gauss Markov (*Gauss Markov Random Field* ou GMRF) sont utilisés dans le système KIM [Datcu et al., 2003]. La dimension fractale [Mandelbrot, 1983], quant à elle, est souvent utilisée pour mesurer la rugosité d'une surface : plus grande est la dimension fractale, plus rugueuse est la texture.

Par ailleurs, certains travaux font une utilisation jointe de la couleur et de la texture. Les auteurs de [Greenhill et al., 2003] par exemple, proposent de calculer des caractéristiques de texture, non pas sur les données d'intensité brutes, mais sur un canal radiométrique transformé, comme le NDVI par exemple. Cependant, dans leur étude comparative des algorithmes de classification des images naturelles de textures en couleurs, Mäenpää et Pietikäinen [Mäenpää & Pietikäinen, 2004] montrent que la couleur et la texture sont des phénomènes séparés qui doivent être traités séparément.

Dans la suite de cette section, nous détaillons quelques méthodes d'extraction des caractéristiques de texture, ayant déjà été mises en oeuvre lors de précédents travaux effectués dans le cadre du Centre de Compétences, et donc particulièrement adaptées à notre problématique.

Caractéristiques de Haralick La méthode de Haralick, présentée dans [Haralick et al., 1973], est basée sur les matrices de cooccurrence. L'idée principale de cette méthode est que toutes les informations de texture peuvent être exprimées par un ensemble de matrices de dépendance spatiale des niveaux de gris, calculées pour différents angles θ , en général $\theta = 0^\circ, 45^\circ, 90^\circ$ et 135° . Soit une image $I(x, y)$, de taille $N_x \times N_y$, ces matrices, dites de co-occurrence, sont calculées pour une paire de pixels de coordonnées (m, n) et $(m \pm \delta, n \pm \delta)$, qui sont séparés de δ pixels et ayant un angle θ par rapport à l'axe horizontal :

$$P(l_i, l_j, \delta, \theta) = \#\{I(m, n) = l_i, I(m \pm \delta, n \pm \delta) = l_j, \theta\} \quad (3.5)$$

où $l_i, l_j \in \{1, 2, \dots, N_g\}$, N_g étant le nombre de niveaux de gris dans l'image.

Partant de l'hypothèse que toutes les informations de texture sont contenues dans les matrices de cooccurrence, 13 descripteurs statistiques en sont extraits. Ces descripteurs, détaillés en annexe B, résument bien le comportement de ces matrices, car ils expriment leur homogénéité, leur contraste, leur corrélation, etc.

Pour une distance δ fixée et 4 directions ($\theta = 0^\circ, 45^\circ, 90^\circ$ et 135°), les caractéristiques statistiques, ainsi que leurs moyennes et variances sont utilisées. Au total, il y a $6 \times 13 = 78$ descripteurs de Haralick pour chaque image.

Caractéristiques de Gabor La méthode Gabor [Dunn & Higgins, 1995; Dunn et al., 1994], est basée sur les filtres de Gabor, qui sont une variante des ondelettes, en termes d'analyse. Les filtres de Gabor simulent bien le système visuel humain (SVH), qui effectue une analyse spatio-fréquentielle minimisant l'incertitude à la fois dans les domaines spatial et fréquentiel.

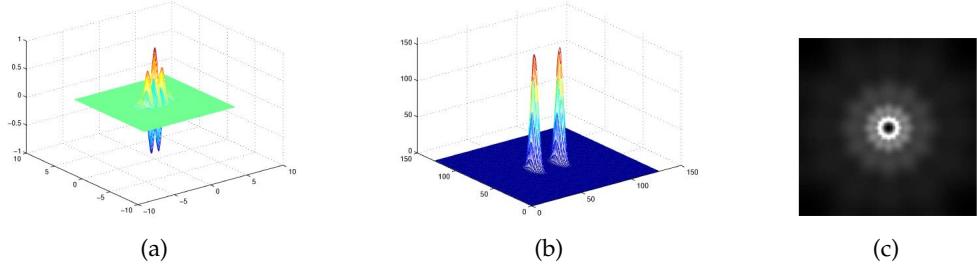


FIG. 3.5 – Exemple d'un filtre de Gabor (a) dans le domaine spatial et (b) dans le domaine fréquentiel. (c) Filtres de Gabor dans le domaine fréquentiel, pour 4 échelles (de 0.05 à 0.4) et 6 orientations (de 0° à 150° par pas de 30°).

Un filtre de Gabor est une fonction sinusoïdale qu'on multiplie par une enveloppe gaussienne, la fonction sinusoïdale étant caractérisée par sa fréquence et par son orientation. Sa réponse dans le domaine spatial est donnée par :

$$h(x, y; u, \theta) = \exp\left(-\frac{1}{2}\left[\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right]\right) \cos(2\pi ux') \quad (3.6)$$

où $x' = x \cos \theta + y \sin \theta$ et $y' = -x \sin \theta + y \cos \theta$. u est la fréquence sinusoïdale le long de la direction θ par rapport à l'axe des x . Les variances σ_x et σ_y spécifient l'enveloppe de la fonction gaussienne et déterminent la largeur de bande du filtre de Gabor.

La transformée de Fourier de l'équation 3.6 est définie comme suit :

$$H(U, V) = 2\pi\sigma_x\sigma_y \left(\exp\left\{-\frac{1}{2}\left[\frac{(U-u)^2}{\sigma_u^2} + \frac{V^2}{\sigma_v^2}\right]\right\} + \exp\left\{-\frac{1}{2}\left[\frac{(U+u)^2}{\sigma_u^2} + \frac{V^2}{\sigma_v^2}\right]\right\} \right) \quad (3.7)$$

avec $\sigma_u = \frac{1}{2\pi\sigma_x}$ et $\sigma_v = \frac{1}{2\pi\sigma_y}$.

La figure 3.5 montre un exemple de filtre de Gabor dans les domaines spatial et fréquentiel.

Les ondelettes de Gabor sont obtenues par translation et dilatation de l'équation 3.6, de telle sorte qu'ils puissent couvrir le domaine fréquentiel quasi uniformément. Ainsi pour une échelle n et une orientation m :

$$g_{mn}(x, y) = a^{-m}h(x', y'), \quad a > 1 \quad (3.8)$$

où

$$x' = a^{-m}(x \cos \theta + y \sin \theta) \quad (3.9)$$

$$y' = a^{-m}(-x \sin \theta + y \cos \theta) \quad (3.10)$$

$$\theta = \frac{n\pi}{K} \quad (3.11)$$

K étant le nombre total d'orientations et a^{-m} est le facteur de la $m^{\text{ième}}$ échelle.

S étant le nombre total d'échelles, les caractéristiques extraites des filtres de Gabor sont la moyenne et la variance des sorties de chacun des $S \times K$ filtres (images filtrées), représentant la statistique de la distribution d'énergie dans chaque filtre. Ces caractéristiques sont calculées dans le domaine fréquentiel, et il est alors possible d'obtenir un vecteur de caractéristiques invariant par rotation [Manthalkar et al., 2003].

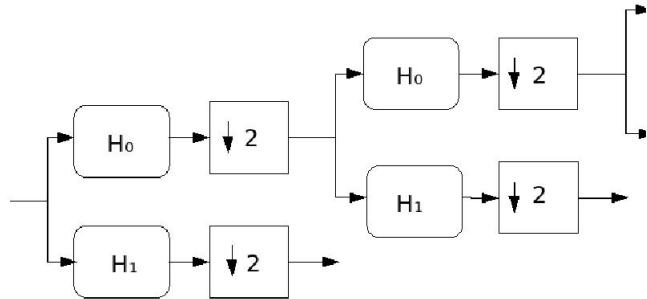


FIG. 3.6 – Décomposition QMF dans le cas monodimensionnel.

Caractéristiques QMF La méthode QMF ou Quadratic Mirror Filters [Mallat, 2003] est basée sur les filtres miroirs en quadrature. Ces derniers sont utilisés pour reconstruire des signaux à partir de composantes basses fréquences et hautes fréquences, bien que ces composantes ne respectent pas le théorème de Shannon : le sous-échantillonnage induit un repliement spectral. Toutefois la forme complémentaire des filtres passe-bas et passe-haut garantit une reconstruction exacte [Vetterli, 1986]. La représentation en ondelettes peut être obtenue par des itérations d'un banc de filtres sur la sortie du canal passe-bas.

Le principe du filtrage QMF est illustré dans la figure 3.6, dans le cas monodimensionnel. Il consiste en deux filtres orthogonaux : un filtre passe-bas H_0 , et un filtre passe-haut H_1 , qui sont appliqués sur le signal d'entrée. Après l'étape de sous-échantillonnage, deux sous-bandes sont obtenues : la sous-bande L correspondant au filtrage passe-bas, et la sous-bande H pour le filtrage passe-haut. L'itération suivante est opérée sur la sous-bande L .

Dans le cas où le signal d'entrée est bidimensionnel, comme les images, le filtrage est effectué dans les directions horizontale et verticale par des filtres séparables, construits par convolution des filtres monodimensionnels. On obtient ainsi 4 sorties : LL , LH , HL et HH , la sous-bande HL par exemple, étant obtenue en appliquant le filtrage passe-haut dans la direction horizontale et le filtrage passe-bas dans la direction verticale. Comme dans le cas monodimensionnel, la décomposition suivante est appliquée sur la sous-bande LL , considérée comme étant la bande fréquentielle la plus représentative.

Les caractéristiques QMF sont alors les moyennes et variances des coefficients des sous-bandes HH , HL et LH , représentant la statistique de la distribution d'énergie dans chaque sous-bande. Pour s échelles, on a donc $s \times 2 \times 3$ caractéristiques QMF pour l'image.

3.2.3 Primitives géométriques

En ce qui concerne les régions issues d'une segmentation ou d'un processus de classification, outre la couleur et la texture, les attributs géométriques sont très souvent utilisés. La géométrie est une caractéristique visuelle importante : elle fait partie des caractéristiques de base, nécessaires pour la description du contenu d'une image.

Zhang et Lu [Zhang & Lu, 2004] et Loncaric [Loncaric, 1998] proposent une analyse étendue des techniques de description de la géométrie. En effet, les auteurs les classifient en descripteurs basés sur les contours et ceux basés sur les régions. L'analyse est détaillée, depuis les caractéristiques simples comme la surface, l'écéntricité, la compacité, les moments, l'orientation, etc, jusqu'aux caractéristiques structurales complexes basées sur la

grammaire. Très souvent, parmi les caractéristiques simples, la quantité $IM = \frac{4\pi S}{P^2}$, appelée indice de Miller, est utilisée pour juger de la régularité de la forme de la région. S et P sont respectivement l'aire et le périmètre de la région. IM est compris entre 0 et 1, et est maximal pour un cercle.

En outre, certains travaux utilisent des modèles de forme pour améliorer les tâches de segmentation [Fua & Hanson, 1987]. Ces modèles sont définis en utilisant une grammaire de primitives de l'image (bords, pixels, ...) et de relations (lignes, coin, jonction T, parallèle, ...). Les travaux de Aksoy [Aksoy et al., 2005] rejoignent ce type de modélisation. Dans [Jibrini, 2002], se trouve une approche plus récente de grammaires de forme pour la reconnaissance de bâtiments dans les images aériennes. Dans [Peura & Iivarinen, 1997], les auteurs ont examiné l'efficacité de quelques descripteurs géométriques simples (convexité, rapport des axes principaux, ...) et affirment qu'en général, il n'est pas nécessaire d'utiliser des descripteurs géométriques complexes et longs à calculer.

Par ailleurs, dans [Pesaresi & Benediktsson, 2001], les auteurs proposent d'utiliser des caractéristiques morphologiques (appelées *Differential Morphological Profiles* ou DMP) pour décrire des objets et structures d'une image à haute résolution. Ainsi, un objet ou une structure de l'image pourrait être une région de pixels avec les mêmes caractéristiques morphologiques. Il est commun d'utiliser les opérateurs morphologiques d'ouverture et de fermeture pour isoler les structures claires et sombres de l'image. Une combinaison de ces deux opérateurs peut donc aboutir à la définition des attributs morphologiques. [Shyu et al., 2007] utilise une extension des DMPs basée sur une approche multi-échelle, pour encoder l'information sur la présence d'objets à différentes échelles spatiales (taille de l'élément structurant).

3.3 Sélection de primitives

Après la procédure d'extraction de primitives, l'image est donc représentée comme un ensemble de ces primitives. Comme indiqué dans la section 2.2.2, ces caractéristiques sont souvent combinées ou concaténées pour améliorer les performances de la classification. Les images peuvent par conséquent être décrites par des dizaines de caractéristiques. Il est donc important de prendre en compte la dimensionnalité des données, c'est-à-dire le nombre de caractéristiques, ce dernier pouvant influencer de manière significative les résultats de classification [Bishop, 2006]. En effet, contrairement à ce que l'intuition peut laisser penser, la performance des classificateurs n'augmente pas indéfiniment avec la taille du vecteur de caractéristiques. Après avoir augmenté en fonction du nombre de primitives, l'efficacité atteint un plateau et quelquefois se met à décroître au-delà d'un certain nombre de primitives : c'est ce qu'on appelle la "malédiction de la dimensionnalité" [Bellman, 1961]. Par ailleurs la complexité de la classification, en termes de temps de calcul, augmente avec la taille du vecteur de caractéristiques. Il est donc intéressant de limiter le nombre de descripteurs effectifs au nombre "optimal", en sélectionnant les descripteurs les plus pertinents : c'est l'objet de la sélection automatique de caractéristiques [Blum & Langley, 1997; Guyon & Elisseeff, 2003].

Afin de déterminer les bonnes caractéristiques à employer pour décrire le contenu des images, plusieurs méthodes et algorithmes supervisés ou non, ont été proposés dans la littérature. Ils se répartissent en deux groupes principaux : les "filters", qui exploitent les propriétés intrinsèques des caractéristiques utilisées sans référence à une quelconque application, et les "wrappers", qui définissent la pertinence des caractéristiques par l'in-

termédiaire de la prédiction de la performance du système final. Campedel et Moulines [Campedel & Moulines, 2005] comparent l'efficacité de tels algorithmes récents de sélection des modèles de texture.

Par ailleurs, dans [Peng et al., 2005] les auteurs proposent un algorithme de sélection d'attributs, à l'aide du critère de dépendance statistique maximale basé sur l'information mutuelle. En utilisant d'abord le critère de redondance minimale et pertinence maximale (mRMR) qu'ils combinent ensuite à d'autres algorithmes de sélection de type "wrapper", ils arrivent à sélectionner à un très faible coût, un ensemble réduit de caractéristiques. Cette méthode, comparée à d'autres algorithmes de sélection tels que la pertinence maximale, en utilisant trois classificateurs différents (le classificateur bayésien naïf, les SVM et l'analyse discriminante linéaire), améliore les performances de classification de manière prometteuse. De même, Charou et ses collègues [Charou et al., 2005] proposent une méthode de sélection d'attributs utilisé en classification d'images multispectrales pour le contrôle de la qualité de l'eau. Ce nouvel algorithme de sélection, nommé *Greedy Non-Redundant* (GreeNRed), est basé sur la théorie de l'information, et sélectionne de manière gloutonne, les caractéristiques ne contenant pas d'information redondante. L'évaluation de l'algorithme de sélection, opérée en utilisant les classificateurs k -plus proches voisins et SVM, montre que GreeNRed permet d'obtenir de meilleures performances de classification.

D'autre part, une méthode de sélection automatique non supervisée, d'attributs issus d'un ensemble très redondant est proposée dans [Campedel et al., 2008]. Cette méthode repose sur l'algorithme de clustering k -moyennes, et sur la sélection des caractéristiques les plus proches des centroïdes des clusters. Les auteurs montrent que, testée avec un classificateur SVM à noyau gaussien sur des images satellitaires SPOT5, la méthode proposée permet d'avoir des performances de classification comparables à celles obtenues en utilisant une méthode supervisée de sélection d'attributs. Tenant compte de ces résultats, un algorithme de sélection basé sur un consensus de clusterings est présenté en outre. Ce dernier combine les résultats des clusterings des k -moyennes, avec différentes initialisations et plusieurs valeurs possibles pour k , afin d'obtenir entre autres, les caractéristiques stables.

3.4 Méthodes de classification supervisée

Nous présentons ci-dessous, des méthodes de classification par apprentissage supervisé. Nous nous intéressons particulièrement à trois d'entre elles, très utilisées dans la littérature pour le problème de la classification d'images : les k -plus proches voisins, la classification bayésienne, et les machines à vecteurs de support (SVM).

3.4.1 k -plus-proches voisins (k -ppv)

L'algorithme des plus proches voisins est l'un des plus simples en apprentissage automatique. Le principe de cette méthode consiste à attribuer à un nouvel objet, la classe majoritaire parmi les classes de ses k plus proches voisins dans l'échantillon d'apprentissage. La notion de voisinage ou de similarité est caractérisée par une mesure de distance dans l'espace des attributs des objets : norme L1, distance euclidienne, distance de Mahalanobis, etc. L'étape d'apprentissage requiert simplement le stockage de l'échantillon d'apprentissage dans un tableau. Le calcul des distances et le tri des plus proches voisins

sont effectués lors de la phase de prédiction d'un objet inconnu, via le parcours exhaustif de l'échantillon d'apprentissage. Un désavantage de cette méthode est la difficulté de traiter avec précision, des bases de données de grandes dimensions en termes de nombre d'attributs ("malédiction de la dimension").

3.4.2 Classification bayésienne

3.4.2.1 Théorie de la décision bayésienne

La théorie de la décision bayésienne, très développée dans la littérature, est une approche statistique fondamentale au problème de la reconnaissance des formes [Duda et al., 2000; Nadler & Smith, 1993]. Cette approche est basée sur l'hypothèse que le problème de décision peut être spécifié en termes de probabilités. Et sous ces hypothèses, la décision bayésienne peut être considérée comme optimale. En fait, les données proviennent d'un processus qui n'est pas entièrement connu. Ce manque de connaissance est signifié en modélisant le processus comme aléatoire, et en utilisant la théorie des probabilités pour l'analyser.

Principe Soient $\{C_k, k = 1 \dots K\}$, un ensemble de K classes, et \mathbf{x} , un vecteur de caractéristiques. Pour chaque classe C_k , on suppose connaître :

- $P(C_k)$, la probabilité a priori de la classe,
- $P(\mathbf{x} | C_k)$, la densité de probabilité de \mathbf{x} , conditionnée par cette classe.

Alors, la probabilité jointe entre l'échantillon \mathbf{x} et la classe C_k se définit comme suit :

$$\begin{aligned} P(C_k, \mathbf{x}) &= P(\mathbf{x}, C_k) \\ P(C_k|\mathbf{x})P(\mathbf{x}) &= P(\mathbf{x}|C_k)P(C_k) \end{aligned}$$

On en déduit la règle de Bayes, qui exprime la probabilité a posteriori de la classe C_k sachant l'exemple \mathbf{x} :

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \quad (3.12)$$

avec $P(\mathbf{x}) = \sum_{k=1}^K P(\mathbf{x}|C_k)P(C_k)$.

C'est une règle d'apprentissage, car elle indique comment ajuster la probabilité attribuée à la véracité d'une hypothèse lorsque notre état de connaissances change avec l'acquisition des données. $P(\mathbf{x}|C_k)$ est appelé vraisemblance, et $P(\mathbf{x})$ est la probabilité marginale de \mathbf{x} , utilisée comme facteur de normalisation.

Ainsi, la règle de Bayes permet de déterminer la probabilité a posteriori que l'exemple \mathbf{x} appartienne à chacune des classes, et la règle de décision de Bayes permet de rendre optimales, les décisions basées sur cette connaissance. En effet, la théorie de la décision bayésienne établit que la stratégie qui consiste à affecter une nouvelle observation à la classe ayant la plus grande probabilité a posteriori est optimale, c'est à dire génère un plus petit nombre d'erreurs que toute autre stratégie.

Cependant, en pratique, nous ne disposons pas de ces probabilités (vraisemblance, probabilité a priori, probabilité marginale), mais uniquement de données exemples \mathbf{x}_n . Il s'agit donc d'estimer ces probabilités à partir d'un ensemble d'apprentissage donné.

Nous considérons une approche paramétrique, dans laquelle nous supposons que les données exemples sont obtenues à partir d'une distribution d'un modèle connu (gaussien par exemple). L'avantage de l'approche paramétrique est que le modèle est défini par un petit nombre de paramètres (par exemple, la moyenne et la variance pour le modèle gaussien), et une fois que ces paramètres sont estimés à partir des données, la distribution complète est connue.

Estimation statistique des paramètres Les approches d'estimation étudient les conséquences du choix d'un modèle particulier, supposé "vrai". En effet, dans un problème d'estimation, on suppose que le modèle est vrai pour une valeur (inconnue) θ_0 de ses paramètres, et l'on explore les contraintes imposées aux paramètres par les données.

Les paramètres du modèle sont estimés à partir des données, de telle sorte qu'ils expliquent le mieux l'ensemble des observations. La distribution estimée est ensuite utilisée pour la prise de décision. Dans la littérature, on trouve deux principales méthodes pour estimer les paramètres d'une distribution : le maximum de vraisemblance et l'estimation bayésienne [McLachlan & Peel, 2000; Bishop, 2006; Duda et al., 2000].

Maximum de vraisemblance Soit $X = \{\mathbf{x}_n\}_{n=1}^N$, une séquence de N vecteurs \mathbf{x}_n indépendants et identiquement distribués (i.i.d.). Supposons que les exemples \mathbf{x}_n sont les réalisations d'une variable aléatoire x à densité de probabilité connue $p(x | \theta)$, θ étant le vecteur de paramètres inconnus gouvernant la distribution de x :

$$\mathbf{x}_n \sim p(x | \theta) \quad (3.13)$$

Les \mathbf{x}_n étant indépendants, la vraisemblance de l'ensemble X , sachant le paramètre θ est le produit des vraisemblances de chaque exemple :

$$L(\theta | X) \equiv p(X | \theta) = \prod_{n=1}^N p(\mathbf{x}_n | \theta) \quad (3.14)$$

La méthode d'estimation par Maximum de Vraisemblance (MV) cherche le paramètre θ qui maximise la vraisemblance de l'ensemble, c'est-à-dire, la probabilité que la variable aléatoire x génère l'ensemble des données X , notée $L(\theta | X)$. En pratique, il est plus simple d'utiliser la log-vraisemblance $\mathcal{L} \equiv \log(L)$. Le problème d'estimation par maximum de vraisemblance devient alors :

$$\hat{\theta}_{MV} = \arg \max_{\theta} \mathcal{L}(\theta | X) = \arg \max_{\theta} \sum_{n=1}^N \log(p(\mathbf{x}_n | \theta)) \quad (3.15)$$

Il est à noter que la solution des *moindres carrés* est un cas particulier de celle du maximum de vraisemblance, lorsque la distribution est gaussienne. Par ailleurs, le très populaire algorithme d'estimation des paramètres, Espérance - Maximisation (*Expectation - Maximisation* ou EM) [Dempster et al., 1977], est basé sur la méthode du maximum de vraisemblance.

Estimation bayésienne Comme son nom l'indique, cette méthode d'estimation fait un large usage de la règle de Bayes. Un exemple d'estimateur bayésien particulier est l'estimateur du Maximum a Posteriori (MAP), qui est très similaire à celui du maximum

de vraisemblance, mais qui permet d'introduire une connaissance a priori sur les paramètres, en les pondérant avec une distribution a priori $p(\theta)$, via la formule de Bayes :

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta | X) = \arg \max_{\theta} (p(X | \theta) p(\theta)) \quad (3.16)$$

Plus généralement, l'estimation bayésienne considère les paramètres comme des variables aléatoires, avec une distribution a priori connue. Cette dernière est utilisée conjointement avec les données, pour calculer la probabilité a posteriori, suivant la règle de Bayes. Et puisque le calcul n'est plus réduit à la recherche d'un maximum, il est nécessaire de calculer le terme de normalisation, c'est-à-dire, la probabilité marginale de $X : p(X)$.

Une présentation plus détaillée ainsi qu'une comparaison de ces approches sont disponibles dans [Heinrich, 2008; Duda et al., 2000].

En pratique, pour des tâches de traitement d'images par exemple, pour lesquelles on peut disposer de centaines de milliers d'échantillons de très grande dimension, l'estimation bayésienne est coûteuse en temps de calcul, tandis que l'estimateur du maximum de vraisemblance est plus rapide et donne des résultats plus faciles à interpréter. En outre, pour de grands volumes de données, les deux approches donnent des résultats similaires. Donc, dans la suite, nous considérerons plus souvent l'estimateur du maximum de vraisemblance.

3.4.2.2 Modèle bayésien naïf

Soit un vecteur $\mathbf{x} \in X$, caractérisé par D attributs : $\mathbf{x} = (x_1, x_2, \dots, x_D)$. Soit θ , l'ensemble des paramètres du modèle décrivant X . D'après la formule de Bayes, la probabilité a posteriori $p(\theta | \mathbf{x})$ est proportionnelle au produit de la probabilité a priori $p(\mathbf{x})$ et de la vraisemblance $p(\mathbf{x} | \theta)$, ces deux dernières probabilités étant à estimer à partir des données.

Cependant, l'estimation de la vraisemblance est souvent plus difficile que celle de l'a priori, car l'espace des caractéristiques peut être très grand et il faudrait un échantillon X de taille trop importante pour pouvoir estimer convenablement ces quantités :

$$p(\mathbf{x} | \theta) = p(x_1, x_2, \dots, x_D | \theta) \quad (3.17)$$

$$p(\mathbf{x} | \theta) = \prod_{d=1}^D p(x_d | x_{d-1}, \dots, \theta) \quad (3.18)$$

Le modèle naïf de Bayes repose sur une hypothèse très forte, qui stipule que la donnée \mathbf{x} est une conjonction de valeurs d'attributs. En d'autres termes, l'hypothèse de Bayes naïve consiste à supposer que les attributs sont des variables aléatoires indépendantes :

$$p(\mathbf{x} | \theta) = \prod_{d=1}^D p(x_d | \theta) \quad (3.19)$$

et chaque facteur $p(x_d | \theta)$ est estimé à l'aide de l'ensemble d'exemples. Cette hypothèse permet de faire les calculs simplement, et finalement, les résultats obtenus ne sont pas sans intérêt d'un point de vue pratique.

Le classificateur bayésien naïf combine ce modèle avec une règle de décision. Soit $\{C_k, k = 1, \dots, K\}$, un ensemble de K classes. Sous les hypothèses usuelles d'existence

des lois de probabilité, le classificateur bayésien naïf attribue toute nouvelle donnée \mathbf{x} , à la classe $C_{\hat{k}}$ qui maximise sa probabilité a posteriori (règle de décision du Maximum a Posteriori) ou sa vraisemblance lorsque l'a priori est uniforme (règle de décision du Maximum de vraisemblance) :

$$\hat{k} = \arg \max_k p(C_k) p(\mathbf{x} | C_k) = \arg \max_k p(C_k) \prod_{d=1}^D p(x_d | C_k) \quad (3.20)$$

Le classificateur bayésien naïf est efficace et, dans certains domaines, ses résultats sont compétitifs à ceux des meilleures méthodes. Cette efficacité s'applique même dans les domaines où la présupposition d'indépendance des attributs ne s'applique pas tout à fait : le domaine de la classification de documents, en particulier, est un domaine pour lequel les classificateurs bayésiens naïfs sont souvent utilisés avec succès, malgré la dépendance entre les attributs (mots d'un document) [Mitchell, 1997; Witten & Frank, 2005].

3.4.2.3 Modèle de mélange de gaussiennes

Les modèles de mélange sont très utilisés pour représenter les données en classification automatique. Une loi de mélange se définit comme une combinaison linéaire de plusieurs fonctions densité $p_k(x)$:

$$\sum_{k=1}^K \alpha_k p_k(x), \quad K > 1 \quad (3.21)$$

avec $\sum_k \alpha_k = 1$, α_k représentant la proportion du modèle k dans le mélange.

En général, les fonctions densité p_k appartiennent à une famille paramétrique, de paramètre inconnu θ_k , conduisant ainsi au modèle de mélange paramétrique :

$$\sum_{k=1}^K \alpha_k p(\mathbf{x} | \theta_k) \quad (3.22)$$

Ainsi, $\alpha_k = p(\mathbf{x} \in k | \theta_k)$, est la probabilité a priori que toute donnée \mathbf{x} appartienne à une seule composante k du mélange, et $p(\mathbf{x} | \theta_k)$ est la densité de probabilité conditionnelle que le vecteur \mathbf{x} appartienne au modèle k .

Dans le cas particulier et très courant où les $p(\mathbf{x} | \theta_k)$ sont des distributions gaussiennes, on parle de mélange gaussien ou *Gaussian Mixture Model* (GMM). θ_k représente alors la moyenne μ_k et la matrice de covariance Σ_k à estimer.

Soit X , un ensemble de N exemples \mathbf{x}_n , ayant chacun D caractéristiques. Sous l'hypothèse que les exemples \mathbf{x}_n , $n = 1 \dots N$ sont indépendamment distribués, X est modélisé par un mélange fini de gaussiennes, de paramètre global Θ et de fonction de vraisemblance $p(X | \Theta)$, définie comme le produit des probabilités conditionnelles de chaque vecteur de caractéristiques \mathbf{x}_n :

$$p(X | \Theta) = \prod_{n=1}^N \sum_{k=1}^K \alpha_k p(\mathbf{x}_n | \mu_k, \Sigma_k) \quad (3.23)$$

où $\Theta = \{\alpha_k, \mu_k, \Sigma_k, k = 1 \dots K\}$ regroupe tous les paramètres des modèles, et $p(\mathbf{x}_n | \mu_k, \Sigma_k)$ est la distribution gaussienne multidimensionnelle :

$$p(\mathbf{x}_n \mid \mu_k, \Sigma_k) = \mathcal{N}(\mathbf{x}_n; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \mu_k)\Sigma_k^{-1}(\mathbf{x}_n - \mu_k)'\right) \quad (3.24)$$

L'estimation des paramètres α_k , μ_k et Σ_k est effectuée par l'algorithme EM basé sur le maximum de vraisemblance et décrit ci-dessous.

Algorithme EM Pour les calculs de maximum de vraisemblance, il est possible d'utiliser des procédures d'optimisation telles que l'algorithme d'Espérance - Maximisation (EM) [Dempster et al., 1977; McLachlan & Peel, 2000]. Il est utilisé pour simplifier les problèmes difficiles de maximum de vraisemblance, en élargissant l'ensemble des exemples avec des données dites latentes ou cachées, dont l'observation aurait simplifié la maximisation de la vraisemblance. Cependant, il est à noter que l'algorithme EM peut ne pas converger vers le principal mode de la vraisemblance. Et dans plusieurs cas, la vraisemblance n'est pas bornée, le résultat de l'estimateur de vraisemblance est alors un maximum local.

L'algorithme EM est un algorithme d'optimisation itératif, chaque itération se résument en deux étapes : l'étape E (Espérance) qui calcule l'espérance de la vraisemblance en incluant les variables latentes comme si elle avaient été observées, et l'étape M (Maximisation) qui consiste à calculer le vecteur de paramètres, en maximisant la vraisemblance trouvée à l'étape E.

Dans le cas particulier d'un modèle de mélange de gaussiennes, la log-vraisemblance de l'ensemble des données X est définie en utilisant les équations 3.23 et 3.15 :

$$\mathcal{L}(\Theta \mid X) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \alpha_k p(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right) \quad (3.25)$$

Cependant, la maximisation de cette fonction par rapport à Θ est très complexe. La méthode de l'algorithme EM consiste donc à ajouter aux données, un complément Z constituant les données manquantes ou cachées. Le couple (X, Z) forme ainsi les données complétées, dont la log-vraisemblance s'exprime par :

$$\mathcal{L}^c(\Theta \mid X, Z) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log (\alpha_k p(\mathbf{x}_n \mid \mu_k, \Sigma_k)) \quad (3.26)$$

où $z_{nk} = 1$ si la donnée \mathbf{x}_n appartient à la composante k du mélange de gaussiennes, et 0 sinon. On considère la loi de Z sachant X , en calculant l'espérance de la vraisemblance complétée, conditionnellement aux données observées et au paramètre courant, noté $\Theta^{(0)}$:

$$\mathcal{Q}(\Theta \mid \Theta^{(0)}, X) = E(\mathcal{L}^c(\Theta \mid X, Z) \mid \Theta^{(0)}, X) \quad (3.27)$$

$$= \sum_{n=1}^N \sum_{k=1}^K E(z_{nk} \mid \Theta^{(0)}, X) \log (\alpha_k p(\mathbf{x}_n \mid \mu_k, \Sigma_k)) \quad (3.28)$$

L'algorithme EM construit donc une suite d'estimateurs $\hat{\Theta}^{(q)}$, $q = 1, 2, \dots$, par itération des étapes Espérance et Maximisation, jusqu'à convergence :

1. **Etape E** : on calcule $\mathcal{Q}(\Theta \mid \Theta^{(q)}, X) = E(\mathcal{L}^c(\Theta \mid X, Z) \mid \Theta^{(q)}, X)$. Si on note $t_{nk}^{(q)} = E(z_{nk} \mid \Theta^{(q)}, X)$, les probabilités a posteriori que \mathbf{x}_n appartienne au modèle k , cette étape nécessite simplement de calculer les $t_{nk}^{(q)}$, en utilisant la loi d'inversion de Bayes :

$$t_{nk}^{(q)} = \frac{\alpha_k^{(q)} p(\mathbf{x}_n \mid \mu_k^{(q)}, \Sigma_k^{(q)})}{\sum_{k=1}^K \alpha_k^{(q)} p(\mathbf{x}_n \mid \mu_k^{(q)}, \Sigma_k^{(q)})} \quad (3.29)$$

2. **Etape M** : on calcule le paramètre $\Theta^{(q+1)}$ qui maximise l'espérance $\mathcal{Q}(\Theta \mid \Theta^{(q)}, X)$

$$\hat{\Theta}^{(q+1)} = \arg \max_{\theta} \mathcal{Q}(\Theta \mid \Theta^{(q)}, X) \quad (3.30)$$

Cette maximisation fournit les paramètres :

$$\alpha_k^{(q+1)} = \frac{1}{N} \sum_{n=1}^N t_{nk}^{(q)} \quad (3.31)$$

$$\mu_k^{(q+1)} = \frac{\sum_{n=1}^N t_{nk}^{(q)} \mathbf{x}_n}{\sum_{n=1}^N t_{nk}^{(q)}} \quad (3.32)$$

$$\Sigma_k^{(q+1)} = \frac{\sum_{n=1}^N t_{nk}^{(q)} (\mathbf{x}_n - \mu_k^{(q+1)})' (\mathbf{x}_n - \mu_k^{(q+1)})}{\sum_{n=1}^N t_{nk}^{(q)}} \quad (3.33)$$

Une propriété de cet algorithme est que la vraisemblance observée augmente à chaque itération : $\mathcal{L}(\hat{\Theta}^{(q+1)} \mid X) \geq \mathcal{L}(\hat{\Theta}^{(q)} \mid X)$. Ainsi, la suite $\hat{\Theta}^{(q)}$ converge vers un extremum local ou un point-selle de la fonction de vraisemblance. Et dans le cas général, il n'y a pas de convergence vers l'extremum global. On peut en déduire que l'algorithme EM dépend de la condition initiale $\theta^{(0)}$. Les versions stochastiques de EM ou *stochastic EM* (SEM) sont des propositions de solutions pour pallier ce problème [Celeux & Diebolt, 1985]. Par ailleurs, une autre difficulté avec l'algorithme EM est que le calcul de $\mathcal{Q}(\Theta \mid \Theta^{(q)}, X)$ n'est pas toujours possible. On peut alors remplacer l'espérance par une approximation de Monte Carlo (MCEM).

D'autres variantes de l'algorithme EM existent, entre autres, la *classification EM* (CEM), qui permet de prendre en compte l'aspect classification lors de l'estimation. [Bishop, 2006; McLachlan & Peel, 2000; Govaert & Nadif, 2005] proposent une présentation détaillée de ces algorithmes dérivés de EM.

3.4.3 Machines à vecteurs de support (SVM)

Les méthodes SVM [Cortes & Vapnik, 1995] ont d'ores et déjà montré leur efficacité dans des problèmes divers d'apprentissages à partir d'exemples pour la classification et la régression. Les classificateurs SVM sont initialement des classificateurs linéaires à deux classes, reposant sur un critère de maximisation de la marge de séparation des deux classes. En effet, il s'agit de déterminer un hyperplan dont la distance aux exemples d'apprentissage les plus proches (vecteurs de support) est maximale. L'hyperplan qui maximise cette distance appelée "marge", est l'hyperplan séparateur optimal (figure 3.7). La robustesse des classificateurs SVM provient de l'introduction d'une pénalisation des erreurs de classification dans le critère à optimiser, pénalisation qui assure une meilleure capacité de généralisation.

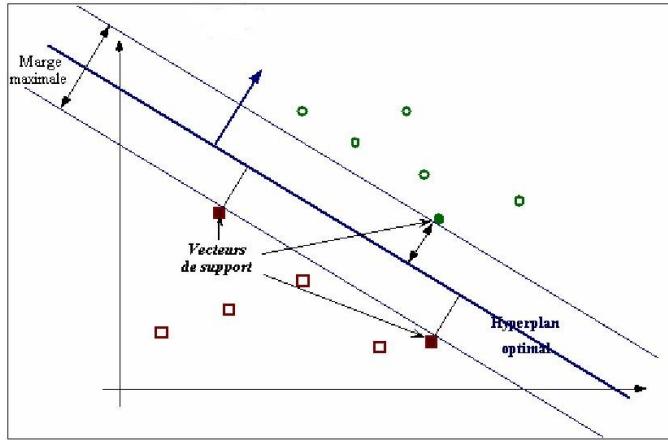


FIG. 3.7 – Principe du SVM : détermination de l’hyperplan optimal maximisant la marge de séparation.

Le problème de classification est le suivant : on souhaite trouver la fonction de décision f , qui basée sur un ensemble X de N observations indépendantes de dimension D , attribue un exemple $x \in X$ à l’une des deux classes $y \in \{-1, +1\}$. La forme générale de la fonction de décision est :

$$f(x) = y = \text{sgn}(g(x)) \quad \text{et} \quad g(x) = \mathbf{w} \cdot x + b \quad (3.34)$$

où ‘.’ représente le produit scalaire, et \mathbf{w} et b , les paramètres à estimer. Le signe de $\mathbf{w} \cdot x + b$ est donc utilisé pour classifier l’exemple x dans l’une des deux classes. En effet, choisir la classe (à valeur dans $\{-1, +1\}$) d’une donnée consiste à déterminer de quel côté du plan d’équation $\mathbf{w} \cdot x + b = 0$ elle se situe.

Lorsque les données d’apprentissage sont linéairement séparables, le classificateur SVM détermine les poids $w_j, j = 1, \dots, D$ et le biais réel b qui maximisent la marge de séparation. Cette dernière étant de $\frac{2}{\|\mathbf{w}\|}$, cela revient à :

$$\begin{aligned} & \text{minimiser} && \|\mathbf{w}\|^2 \\ & \text{sous la contrainte} && y_i(\mathbf{w} \cdot x_i + b) \geq 1, \quad i = 1, \dots, N. \end{aligned}$$

qui exprime la condition de parfaite classification.

Les vecteurs de support sont les données x_i qui vérifient exactement la contrainte : $y_i(\mathbf{w} \cdot x_i + b) = 1$.

On passe au problème dual en introduisant les multiplicateurs de Lagrange pour chaque contrainte. C’est un problème de programmation quadratique de dimension N :

$$\begin{aligned} & \text{maximiser} && G(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ & \text{sous la contrainte} && \sum_{i=1}^N \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, N. \end{aligned}$$

où $K(x_i, x_j) = x_i \cdot x_j$ est la fonction noyau dans le cas linéaire, et les α_i sont tous nuls, exceptés ceux correspondant aux vecteurs de support. La résolution de cette dernière équation permet de déterminer le score du classificateur $g(x)$ et par conséquent, la classe prédite y :

$$g(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b \quad \text{et} \quad y = \text{sgn}(g(x)). \quad (3.35)$$

Lorsque les données ne sont pas linéairement séparables, la fonction noyau K correspond à une transformation Φ qui projette les données dans un espace dans lequel elles seront considérées comme linéairement séparables :

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \quad (3.36)$$

Et dans ce cas, $f : f(x) = \text{sgn}(\mathbf{w} \cdot \Phi(x) + b)$. Cependant, le problème de la classification SVM s'exprime uniquement à l'aide de K , sans avoir à expliciter la fonction Φ . Les noyaux les plus utilisés sont le noyau polynomial et le noyau gaussien qui ne nécessite qu'un paramètre.

Prenant en compte le cas réel où les données sont bruitées, une contrainte de régularisation est introduite. Le problème peut alors prendre la forme suivante :

$$\begin{aligned} & \text{minimiser} && \|\mathbf{w}\|^2 + C \sum_{i=1}^N \varepsilon_i \\ & \text{sous la contrainte} && y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0, \quad i = 1, \dots, N \end{aligned}$$

Nous utilisons dans notre étude, l'implantation des SVM appelée "libsvm" version 2.5 [Chang & Lin, 2004].

3.5 Evaluation de la classification

Il est nécessaire d'évaluer la qualité d'une classification afin de pouvoir comparer des méthodes de classification, et en choisir la meilleure pour une application donnée, ou pour sélectionner le meilleur ensemble de paramètres. En particulier, cette étape est indispensable pour comparer une classification à une classification de référence (carte thématique).

Dans [Smits & Dellepiane, 1999], les auteurs explorent le problème de l'évaluation des algorithmes de classification de l'occupation des terres, et concluent que les méthodes d'évaluation basées sur les matrices de confusion et la statistique KHAT [Wilkinson, 2005] sont les plus appropriées pour comparer les classificateurs.

Encore appelée tableau de contingence, la matrice de confusion est une matrice affichant les statistiques de la précision de classification d'une image, notamment le degré de classification erronée parmi les diverses classes. On peut en extraire la précision globale de la classification, qui est la proportion du nombre total de pixels correctement classifiés par rapport au nombre total de pixels de la matrice. Il est également possible d'en déduire des mesures de la qualité de la classification pour chaque classe : la précision de production et la précision de l'utilisateur. La précision de production donne une mesure de la qualité de l'analyse lors de la production de la carte classifiée, et s'exprime par la proportion des pixels correctement classifiés dans la classe par rapport aux pixels appartenant à cette classe dans la classification de référence. La précision de l'utilisateur, quant à elle, fournit à l'utilisateur une mesure de la probabilité d'une classification adéquate des pixels durant le processus de classification de la carte. Elle s'exprime par le rapport du nombre de pixels correctement classifiés dans la classe sur le nombre total de pixels attribués à cette classe par le système de classification.

La statistique KHAT, approximation du paramètre *Kappa* [Mitra et al., 2002] est une mesure moins intuitive, mais plus objective de la qualité globale de la classification. Elle s'exprime par :

$$KHAT = \frac{N \sum_{i=1}^K X_{ii} - \sum_{i=1}^K (X_{i+} \times X_{+i})}{N^2 - \sum_{i=1}^K (X_{i+} \times X_{+i})} \quad (3.37)$$

où K est le nombre de classes, N est le nombre total d'exemples, X_{ii} est le nombre de pixels correctement classifiés de la classe i , X_{i+} est le nombre de pixels appartenant à la classe i selon la classification de référence, et X_{+i} est le nombre de pixels classifiés par le classificateur dans la classe i .

Meila [Meila, 2002] propose un critère basé sur la théorie de l'information, pour comparer deux partitions d'un même ensemble de données. Ce critère, appelé "variation de l'information (VI)", mesure la quantité d'information perdue et gagnée lorsqu'on passe d'une partition à l'autre.

Par ailleurs, Canters [Canters, 1997], dans ses travaux sur la classification floue de l'occupation des terres, utilise la probabilité a posteriori d'appartenance à chaque classe avec un classificateur basé sur le maximum de vraisemblance, comme une mesure de la qualité de classification de chaque pixel.

3.6 Sélection de modèles

La sélection de modèles est un problème bien connu en statistiques, apprentissage automatique et fouille de données. Etant donné un ensemble d'apprentissage composé de paires entrées / sorties, un modèle est construit pour prédire la sortie à partir de l'entrée. Mais l'apprentissage est un problème mal-posé, et dans un grand nombre de situations, les connaissances a priori sur les données ne permettent pas de déterminer un unique modèle pour réaliser l'inférence. Nous devons donc faire des hypothèses supplémentaires pour trouver une unique solution à partir des données. L'ensemble de ces hypothèses est appelé "biais inductif", sans lequel l'apprentissage est impossible. La sélection de modèles a pour objectif de choisir le "bon" biais [Alpaydin, 2004]. En d'autres termes, c'est la tâche qui consiste à choisir un modèle avec un biais inductif correct, parmi un ensemble de modèles potentiels, ce qui en pratique, signifie sélectionner des paramètres dans le but de créer un modèle de complexité optimale, étant donné l'ensemble d'apprentissage.

Depuis la fin des années 70, des méthodes pour la sélection de modèles à partir de données ont été développées. Les exemples classiques d'application de ces méthodes sont la sélection de variables, ou le choix du nombre de composantes d'un mélange de lois ou de l'ordre d'une chaîne de Markov. Soit X , un ensemble de N données, supposons que nous disposons d'un ensemble de modèles candidats \mathcal{M}_m , $m = 1, \dots, M$ et des paramètres correspondants θ_m , de dimension d_m . Nous souhaitons trouver le modèle décrivant le mieux l'ensemble X . Une taxonomie des méthodes de sélection de modèles les classifie en méthodes empiriques telles que la validation croisée et le rapport signal sur bruit, et en méthodes théoriques comme le *Bayesian Information criterion* (BIC) et le *Minimum Description Length* (MDL).

3.6.1 Validation croisée

En pratique, la validation croisée est l'une des méthodes les plus utilisées pour l'estimation de l'erreur de prédiction. En effet, cette heuristique est facile à utiliser, à implémenter et à comprendre (au moins au niveau intuitif). En outre, elle ne nécessite pas d'hy-

pothèses supplémentaires et donne des résultats qui ont prouvé être difficiles à améliorer en général.

La validation croisée estime directement l'erreur de généralisation. S'il est suffisamment grand, l'ensemble de données est divisé en deux parties : un ensemble d'apprentissage et un ensemble de validation, utilisé pour évaluer la performance du modèle obtenu à partir de l'ensemble d'apprentissage. En général, l'ensemble des données est divisé de manière aléatoire en K parties égales. Ainsi, à chaque boucle, $K - 1$ sous-ensembles sont utilisées pour l'apprentissage du modèle, et l'erreur de prédiction du modèle construit est calculée sur le sous-ensemble restant. Chaque sous-ensemble est utilisé une fois et une seule pour les tests. La procédure est récapitulée dans l'algorithme 3.6.1.

Algorithme 3.6.1 : Procédure de validation croisée

- 1 : L'ensemble X des N données (x_n, y_n) est divisé aléatoirement en K parties égales $X_i, i = 1, \dots, K$.
 - 2 : Pour chaque i , une fonction \hat{f}_i est obtenue à partir des données d'apprentissage $\bigcup_{j \neq i} X_j$. L'erreur de généralisation, calculée sur X_i , est estimée par :

$$CV(\hat{f}_i, X_i) = \frac{1}{|X_i|} \sum_{n=1}^N L(y_n, \hat{f}_i(x_n)),$$
 où $L(y_n, \hat{f}_i(x_n)) = I(y_n \neq \hat{f}_i(x_n))$ par exemple.
 - 3 : L'ensemble X entier est ensuite utilisé comme ensemble d'apprentissage, pour obtenir une fonction \hat{f} . L'estimée de l'erreur de prédiction de la validation croisée $CV(\hat{f})$ est la moyenne des estimées $CV(\hat{f}_i, X_i)$.
-

Etant donné un ensemble de modèles, la validation croisée est utilisée pour la sélection de modèles en calculant l'estimée de l'erreur de prédiction pour chaque modèle candidat, puis en choisissant le modèle qui a la plus petite estimée de l'erreur de prédiction de la validation croisée.

En général, K prend les valeurs 5 ou 10, et lorsque $K = N$, il s'agit de la validation croisée *Leave-one-out*, utilisée en général lorsque N est petit.

Cependant, en pratique, il n'y a pas de garantie que l'estimée de l'erreur de prédiction $CV(\hat{f})$ soit proche de l'erreur. Malgré cela, la sélection de modèles par validation croisée reste très utilisée et a prouvé son efficacité dans une vaste gamme de problèmes.

3.6.2 Les méthodes statistiques

L'une des réponses apportées par les statisticiens au problème de la sélection de modèles est la minimisation d'un critère pénalisé. Les premiers critères apparaissant dans la littérature sont basés sur la théorie de l'information : l'*Akaike Information Criterion* (AIC) [Akaike, 1973], le *Bayesian information Criterion* (BIC) [Schwarz, 1978] et le *Minimum Description Length* (MDL) [Rissanen, 1978]. D'un point de vue théorique, plusieurs travaux ont été réalisés concernant leurs propriétés statistiques et leur adaptation à des modèles spécifiques. En particulier, plusieurs versions corrigées du critère AIC sont été proposées [Hurvich & Tsai, 1989; Sugiura, 1978]. Il existe ainsi une littérature très fournie sur la sélection de modèles par critère pénalisé, qui se développe encore actuellement avec l'utilisation d'outils complexes de probabilité, comme par exemple les inégalités de concentration et de déviation, permettant à la fois la construction des critères et leur

étude.

3.6.2.1 Akaike Information Criterion

Nous souhaitons trouver le modèle qui approxime le mieux la réalité, à partir d'un ensemble de données. En d'autres termes, il s'agit de minimiser la fonction de perte d'information, qui mesure l'erreur entre le modèle et la réalité. L'information de Kullback-Leibler, développée dans [Kullback & Leibler, 1951], est une mesure pour représenter l'information perdue lors de l'approximation de la réalité, sachant qu'un bon modèle minimise la perte d'information. Quelques années plus tard, Akaike [Akaike, 1973] propose d'utiliser cette mesure pour la sélection de modèles, et établit un lien entre la méthode d'estimation du maximum de vraisemblance et l'information de Kullback-Leibler. L'AIC est donc un critère pour estimer l'information de Kullback-Leibler, et est définie comme suit, pour un ensemble de données X et un modèle dont le vecteur de paramètres θ est de dimension d :

$$AIC = -2 \log P(X | \hat{\theta}) + 2d \quad (3.38)$$

Ainsi, pour un ensemble de modèles candidats $\mathcal{M}_m, m = 1, \dots, M$, le meilleur modèle $\mathcal{M}_{\hat{m}}$ est celui qui minimise le critère AIC :

$$\hat{m} = \arg \min_m AIC_m = \arg \min_m \left(-2 \log P(X | \hat{\theta}_m) + 2d_m \right) \quad (3.39)$$

Dans ses travaux, Stone [Stone, 1977] montre que le critère AIC et la validation croisée *leave-one-out* sont asymptotiquement équivalents.

3.6.2.2 Bayesian Information Criterion

Le critère BIC [Schwarz, 1978] se place dans un contexte bayésien de sélection de modèles et, comme le critère AIC, est applicable dans les situations où l'apprentissage est effectué par la maximisation d'une log-vraisemblance. En effet, le critère BIC est une approximation du calcul de la vraisemblance des données conditionnellement au modèle fixé. Il est donc défini par :

$$BIC = -2 \log P(X | \hat{\theta}) + d \log(N) \quad (3.40)$$

pour un ensemble X de N données et un modèle dont le vecteur de paramètres θ est de dimension d . Comme pour le critère AIC, la minimisation de BIC permet d'identifier le meilleur modèle.

Les critères AIC et BIC ont souvent fait l'objet de comparaisons [Burnham & Anderson, 2002; Bozdogan, 1987]. Dans la pratique, le critère BIC a tendance à sélectionner les modèles simples, contrairement à AIC qui choisit des modèles plus complexes. Ceci n'est guère surprenant puisque le critère BIC a un terme de pénalité plus lourd que AIC.

3.6.2.3 Minimum Description Length

L'approche MDL donne un critère de sélection formellement identique au critère BIC, mais établi d'un point de vue de codage optimal, l'apprentissage étant vu comme un problème de compression des données. Introduit par Rissanen [Rissanen, 1978, 1983], le

principe de la minimisation de la longueur de description est basé sur la minimisation de la complexité stochastique.

La complexité stochastique a ses racines dans la notion de complexité algorithmique [Kolmogorov, 1965]. L'idée de base est que le meilleur modèle pour décrire un ensemble de données est celui qui permet de le coder avec la plus petite longueur, exprimée en bits. En d'autres termes, l'ensemble de données est d'autant mieux compris que la longueur du code utilisée pour le représenter est minimale. En effet, toutes les régularités ou dépendances d'un ensemble de données doivent être extraites pour pouvoir compresser au maximum l'information qui y est contenue. Plus la longueur de code est courte, plus l'ensemble de données est régulier ou simple. Le principe de la minimisation de la complexité stochastique peut donc être vu comme une formalisation du principe du "rasoir d'Ockham" [Blumer et al., 1987], qui stipule que le modèle le plus simple expliquant toutes les observations est le meilleur.

Les théorèmes de codage de Shannon donnent le nombre minimal de bits nécessaires pour encoder les données. Cependant, en réalité, pour compléter la description, l'on a aussi besoin de spécifier le modèle parmi un certain nombre de choix, et cela nécessite aussi un certain nombre de bits. La longueur de description L_X d'un ensemble X de N données décrites par un modèle dont le vecteur de paramètres θ est de dimension d , est ainsi constituée de deux parties :

- la longueur de code nécessaire pour décrire les paramètres du modèle : $L(\hat{\theta})$,
- et la longueur de code nécessaire pour décrire les données connaissant les paramètres du modèle : $L(X | \hat{\theta})$.

La longueur de description prend ainsi la forme d'une vraisemblance pénalisée, la pénalité étant le coût de l'encodage des paramètres estimés ($L(\hat{\theta})$). Dans la littérature,

$$L(\hat{\theta}) \approx \frac{d}{2} \log(N) \quad \text{et} \quad L(X | \hat{\theta}) = -\log(P(X | \hat{\theta})) \quad (3.41)$$

Au final :

$$L_X = L(X | \hat{\theta}) + L(\hat{\theta}) \quad (3.42)$$

$$L_X = -\log P(X | \hat{\theta}) + \frac{d}{2} \log(N) \quad (3.43)$$

Etant donné un ensemble de modèles candidats pour représenter un ensemble de données, le principe MDL préconise de choisir le modèle qui minimise l'équation 3.43.

3.7 Méthodes de régularisation

La régularisation est la procédure qui consiste à corriger une image classifiée pour obtenir une structure acceptable : la sortie d'un système de classification est souvent imparfaite, en particulier lorsque la classification s'est déroulée selon une approche pixel. En effet, la classification basée sur le pixel suppose en général l'indépendance d'un pixel vis-à-vis de son environnement, ce qui introduit un bruit "poivre et sel" dans l'image classifiée, qui est généralement supprimé en prenant en compte l'information spatiale et contextuelle.

3.7.1 Vote majoritaire

La post-classification par vote majoritaire permet de prendre en compte le contexte d'un pixel afin d'améliorer la classification finale sur l'image. Ce traitement, opéré sur la

sortie du classificateur, parcourt simplement l'image classifiée avec une fenêtre entourant le pixel classé et détermine la classe majoritaire autour de ce pixel. Si la classe majoritaire est complètement différente du pixel central, on lui attribue une nouvelle classe, sinon on ne fait rien. Cette technique est très utilisée pour éliminer tous les défauts de classification liés au bruit sur les images ou à la mauvaise estimation statistique des paramètres.

3.7.2 Champs de Markov

Une méthode de régularisation très connue utilise les champs de Markov (*Markov Random Fields* ou MRF) [Kindermann & Snell, 1980], initialisé avec le résultat de la classification. Le formalisme des champs de Markov permet aisément d'exprimer des contraintes de voisinage entre les pixels et d'améliorer la classification obtenue [Huang & Liao, 2008; Pony et al., 2000; Jhung & Swain, 1994]. En effet, ce type de méthode suppose en général une dépendance locale entre un pixel et son voisinage, et le résultat de la classification est obtenu de manière itérative.

Une image est modélisée par deux champs aléatoires : $X = \{x_s, s \in S\}$, le champ des classes, et $Y = \{y_s, s \in S\}$, le champ des observations, S représentant l'ensemble des pixels ou sites s de l'image. x_s est une valeur dans un ensemble fini de classes Λ de cardinal K . D'une façon générale, X est un processus contenant l'information que l'on cherche mais n'est pas directement observable. On observe $Y = y$ et on cherche à trouver ou à estimer, la réalisation inaccessible $X = x$, à partir du champ Y .

Le champ des classes X est supposé markovien, autrement dit la probabilité de chaque pixel individuel s ne dépend que de son voisinage V_s . On a alors pour tout s :

$$P(X_s = x_s / x^s) = P(X_s = x_s / x_t, t \in V_s) \quad (3.44)$$

où $x^s = (x_t)_{t \neq s}$ est la configuration de l'image excepté le site s , et V_s est le système de voisinage pour le site s , et est défini comme :

1. $s \notin V_s$
2. et $t \in V_s$ si et seulement si $s \in V_t$.

En supposant que le voisinage et la loi conditionnelle de X_s ne dépendent pas de la position de s et que toutes les configurations X sont de probabilité non nulle (théorème de Hammersley-Clifford), X suit une distribution de Gibbs. Sous cette hypothèse, la probabilité a priori de X s'écrit :

$$P(X = x) = \frac{1}{Z} \exp(-U(x)) = \frac{1}{Z} \exp \left(- \sum_{c \in C} U_c(x) \right) \quad (3.45)$$

où Z est une constante de normalisation incalculable en pratique, $U(x)$ est l'énergie totale de la configuration x , et $U_c(x)$ est le potentiel d'une clique, une clique c étant un ensemble non vide de sites vérifiant : $\forall (s, t) \in c, (s \neq t \Rightarrow s \in V_t)$. La probabilité conditionnelle locale en un site s s'écrit donc :

$$P(X_s = x_s / X^s = x^s) = \frac{\exp(-U_s(x_s / V_s))}{\sum_{x_s \in \Lambda} \exp(-U_s(x_s / V_s))} \quad (3.46)$$

où $U_s(x_s / V_s) = \sum_{c \in C, s \in c} U_c(x_s, V_s)$ est l'énergie locale, ne faisant intervenir que les voisins de s . L'équation 3.46, qui permet donc de calculer la probabilité conditionnelle locale

en chaque site, est à la base de tous les algorithmes de simulation des champs markoviens.

Une méthode de classification contextuelle de type markovien, très présente dans la littérature est l'algorithme des modes conditionnels itérés (*Iterated Conditional Mode* ou ICM).

ICM est un algorithme de relaxation déterministe et itératif, basé sur une stratégie de descente de gradient qui converge rapidement vers un minimum local de la fonction d'énergie. Il peut être considéré comme un cas particulier du recuit simulé, avec une température nulle. Son principe est le suivant : partant d'une configuration initiale d'un champ d'étiquettes, tous les sites de l'image sont visités suivant une stratégie prédéfinie, et leurs étiquettes sont mises à jour par celles qui minimisent l'énergie locale conditionnelle. Puisque la modification de l'étiquette d'un site peut modifier l'énergie locale des sites voisins, le processus est répété jusqu'à convergence vers un minimum local de l'énergie globale. Cet algorithme est très rapide, contrairement au recuit simulé, et peu coûteux en temps de calcul. Cependant ses performances dépendent très fortement de l'initialisation, puisqu'il converge vers un minimum local.

3.8 Conclusions

Dans ce chapitre, nous avons présenté en détail chacune des étapes importantes du processus de classification supervisée des images basée sur une procédure d'apprentissage automatique. En effet, ce sont des notions importantes que nous devons prendre en compte pour espérer mener à bien notre tâche de classification supervisée des images satellitaires. En effet, que ce soit le type de primitives extraites de l'image, l'algorithme d'apprentissage utilisé ou le modèle adopté pour la classification, ou encore la méthode d'évaluation de la qualité des résultats, ils doivent être choisis en fonction du type de données disponibles, mais aussi de l'application. Un compromis est à trouver pour chaque type d'application, entre l'utilisation systématique des traitements simples et rapides et la mise en oeuvre d'algorithmes efficaces, mais de complexité informatique importante.

Chapitre 4

Apprentissage automatique de la production des cartes d'occupation d'occupation du sol

La classification des données de couverture et d'occupation des terres, généralement effectuée par des experts, requiert un effort humain considérable et coûteux, dû au manque de méthodes automatiques pour traiter des tâches complexes de classification telles que l'extraction des classes de CORINE Land Cover (CLC). Les cartes CORINE Land Cover sont générées à partir d'images satellitaires qui sont interprétées visuellement par un photointerpréteur, s'aidant de données exogènes (photographies aériennes, cartes topographiques, etc). Des centaines de milliers de km^2 devant être décrits, il serait donc intéressant de disposer de méthodes automatiques pour la génération et la mise à jour de telles cartes.

Nous nous proposons donc dans ce chapitre, d'étudier les possibilités d'apprentissage automatique de classification d'un système d'interprétation des images pour l'observation de la Terre, par l'utilisation des bases de données constituées par CORINE Land Cover. En effet, nous souhaitons utiliser la partition géographique et la classification de CORINE Land Cover pour guider et renseigner un système automatique d'interprétation d'images, utilisant uniquement des informations contenues dans les images (données spectrales, textures, etc.) pour apprendre des caractéristiques spécifiques des paysages. Ces caractéristiques contribuent à l'indexation des images et permettront de retrouver ces paysages dans d'autres scènes.

Etant donné la forte utilisation des méthodes de classification basées sur le maximum de vraisemblance dans la littérature (voir section 2.2.2), nous proposons dans ce chapitre, d'illustrer une application de l'apprentissage automatique avec de telles méthodes classiques. Nous pourrons ainsi témoigner de la portée de ce type d'approche, et nous aguerrir grâce à ses limitations.

4.1 Description des données

Notre objectif est d'utiliser CORINE Land Cover pour l'apprentissage automatique en télédétection. En effet, nous nous proposons d'utiliser une classification conforme à celle de CORINE Land Cover sur des zones test pour lesquelles nous posséderons :

- les cartes d'occupation des sols conformes aux recommandations de CORINE,
-

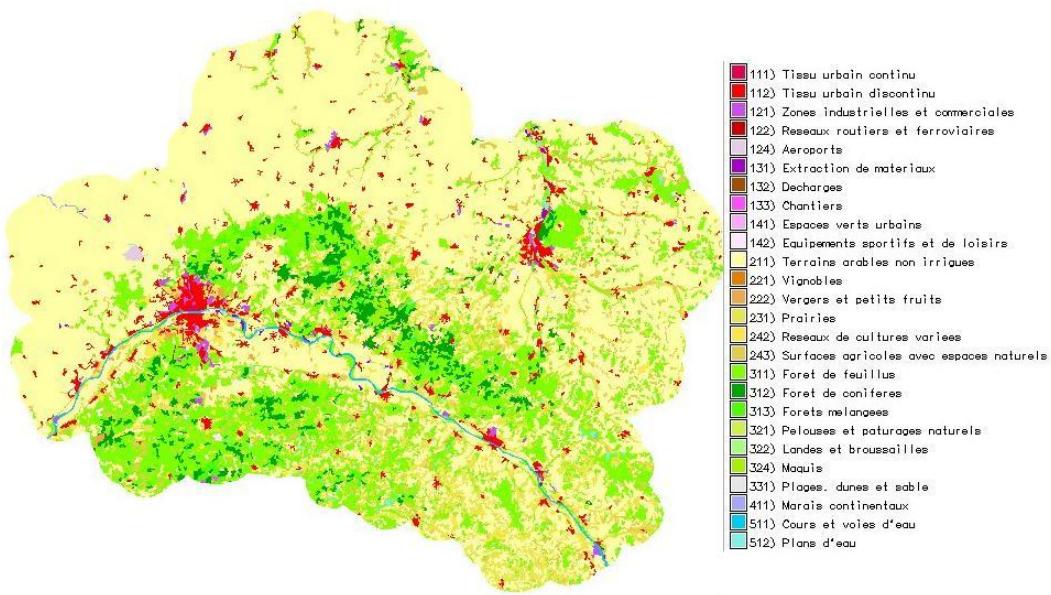


FIG. 4.1 – Carte CORINE Land Cover de la région du Loiret et la légende associée.

– les images satellites multispectrales ayant contribué à la construction de ces cartes,
de façon à apprendre des règles de décision pour la classification automatique de terrains inconnus à partir d'images de télédétection.

4.1.1 CORINE Land Cover

CORINE Land Cover est un inventaire biophysique de l'occupation des terres, fournant une information géographique de référence pour 29 Etats européens et pour les bandes côtières du Maroc et de la Tunisie. Cette description repose sur une taxinomie précise afin d'assurer la consistance des résultats produits par les photointerprètes. Cette base de données géographiques est produite, gérée et utilisée à l'aide d'un Système d'Information Géographique (SIG). La figure 4.1 montre un exemple de carte CLC de la région du Loiret (France), et la légende associée. Une présentation détaillée de CORINE Land Cover est disponible en annexe A, incluant la définition des principes de base à respecter lors de la génération des cartes, un descriptif de la méthode utilisée par les photointerprètes (section A.1), ainsi que la nomenclature standard hiérarchisée en 3 niveaux et 44 postes (section A.2). Néanmoins, afin que le lecteur prenne pleinement conscience du problème, nous souhaitons insister ici sur certains aspects qui traduisent la complexité de cette classification hiérarchique, et donc la délicatesse de la tâche d'automatisation du processus de génération des cartes CORINE Land Cover.

Tout d'abord, l'unité spatiale au sens de CLC est une zone dont la couverture peut être considérée comme homogène, ou être perçue comme une combinaison de zones élémentaires qui représente une structure d'occupation, le critère d'homogénéité selon CLC stipulant qu'au moins 75% de la superficie de l'unité paysagère doit appartenir à une même classe d'occupation du sol. Ceci signifie que dans une région appartenant à

une certaine classe CLC, 25% des terres peuvent appartenir à une autre classe, pouvant ainsi entraîner certaines confusions.

De plus, la légende de CLC ne fait pas uniquement référence à l'occupation du sol, mais aussi parfois à son utilisation. Par conséquent, la classification hiérarchique CORINE Land Cover propose aussi des classes qui sont plus du type *land use* que *land cover*. C'est le cas des classes du type 1 d'occupation du territoire (*Territoires artificialisés*), dans lesquelles pour chacune des classes, on retrouve un peu de bâti, des espaces verts, des routes, ... mais dans des proportions différentes. L'identification automatique de ces classes qui font référence à un usage des sols peut se heurter à des difficultés particulières.

En outre, le respect du critère de superficie minimale (25 hectares) a souvent poussé à regrouper certains modes d'occupation du territoire au sein de postes appelés "postes à caractère mixte" (classes 242 et 243). En effet, ce sont des "pots-pourris" de petites régions de catégories différentes, mais mises ensemble pour satisfaire la condition de superficie minimale de l'unité cartographiée. Par ailleurs, lorsqu'une zone homogène est inférieure à 25 hectares, des règles précises de généralisation ont été adoptées, ces dernières variant suivant la classe concernée.

La constitution des cartes CLC exige donc de respecter un certain nombre de règles qui doivent toutes être prises en compte lors de la génération automatique de telles cartes.

Pour notre étude, nous disposons des données CLC pour les régions du Loiret (45), du Bas-Rhin (67) et de la Somme (80). Ces cartes sont issus de la base CLC90, réalisée à partir d'images satellitaires acquises entre 1987 et 1994. La figure 4.1 montre la carte CLC de la région du Loiret, accompagnée de sa légende.

4.1.2 Images SPOT

Nous avons à notre disposition, des scènes SPOT à 20 m de résolution et 3 canaux : rouge (R), vert (V), et Proche InfraRouge (PIR). Ces images SPOT correspondent à certaines régions du Loiret, du Bas-Rhin et de la Somme pour lesquelles nous possédons des cartes CLC¹. Les caractéristiques des 5 scènes SPOT reçues sont récapitulées dans le tableau 4.1.

TAB. 4.1 – Caractéristiques des données images SPOT.

Région	Scènes SPOT	Format	Date de prise de vue	Corrections
Loiret	39-253	TIFF	08-09-1989	radiométrique, géoréférencée
	41-253	TIFF	09-09-1989	radiométrique, géoréférencée
Bas-Rhin	50-251	Brut	12-07-1990	radiométrique, non géoréférencée
Somme	37-248	TIFF	13-10-1992	radiométrique, géoréférencée
	40-249	Brut	30-07-1992	radiométrique, non géoréférencée

Nous avons par ailleurs acquis, pour ces régions, les cartes numériques de l'IGN (références CDB367N, CDB380E, CDB380O, CDB345E, CDB345O), qui nous donnent des informations complémentaires sur la topographie et l'occupation des sols.

¹Les cartes CLC issues de l'IFEN, ainsi que les images SPOT nous ont été fournies par le CNES.

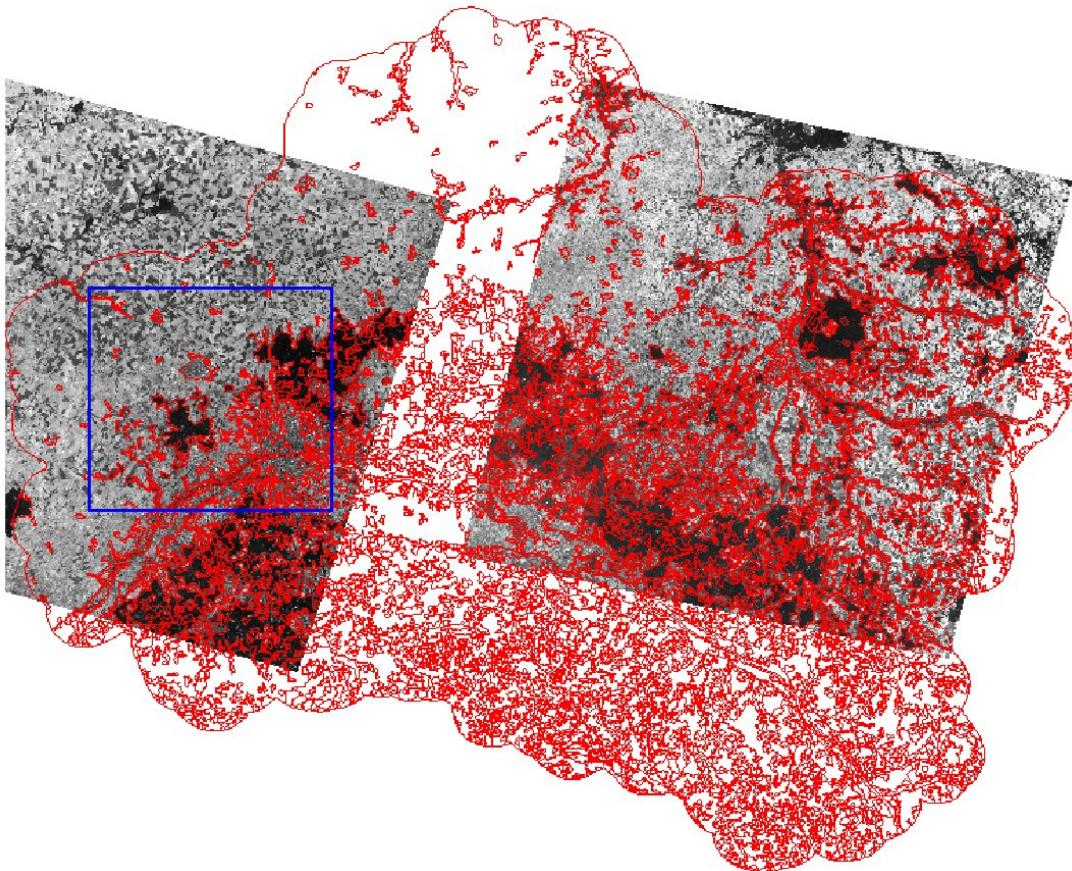


FIG. 4.2 – Couche CLC vectorielle de la région du Loiret superposée à deux scènes SPOT (39-253 à gauche et 41-253 à droite). L'apprentissage de la classification ne pourra s'effectuer que sur les zones communes (intersections). Le rectangle bleu représente la zone utilisée pour les tests.

4.1.3 Gestion des données

Les données ont été manipulées grâce au système d'information géographique (SIG) GRASS².

Pour stocker les données et les cartes, il existe principalement deux formats : le format raster ou point par point, dans lequel les cartes sont en fait un ensemble de points d'une grille régulière et dense auxquels on peut rattacher des données comme l'altitude par exemple, et le format vectoriel dans lequel on définit des zones géographiques à l'aide de vecteurs, les données ne sont plus rattachées à des points mais à des polygones.

Les cartes CLC sont disponibles au format vectoriel et les images satellitaires, au format raster. Ainsi, avec GRASS, il a été possible de :

- visualiser les données CLC vecteur puis raster de nos 3 régions, et d'y associer le code des couleurs des cartes CLC.
- importer les images SPOT dans GRASS et les superposer avec les couches CLC

²GRASS : Geographic Resources Analysis Support System, est un SIG open source, sous licence GPL, utilisé pour le traitement des images, la production de graphiques, la modélisation spatiale et la visualisation de plusieurs types de données.

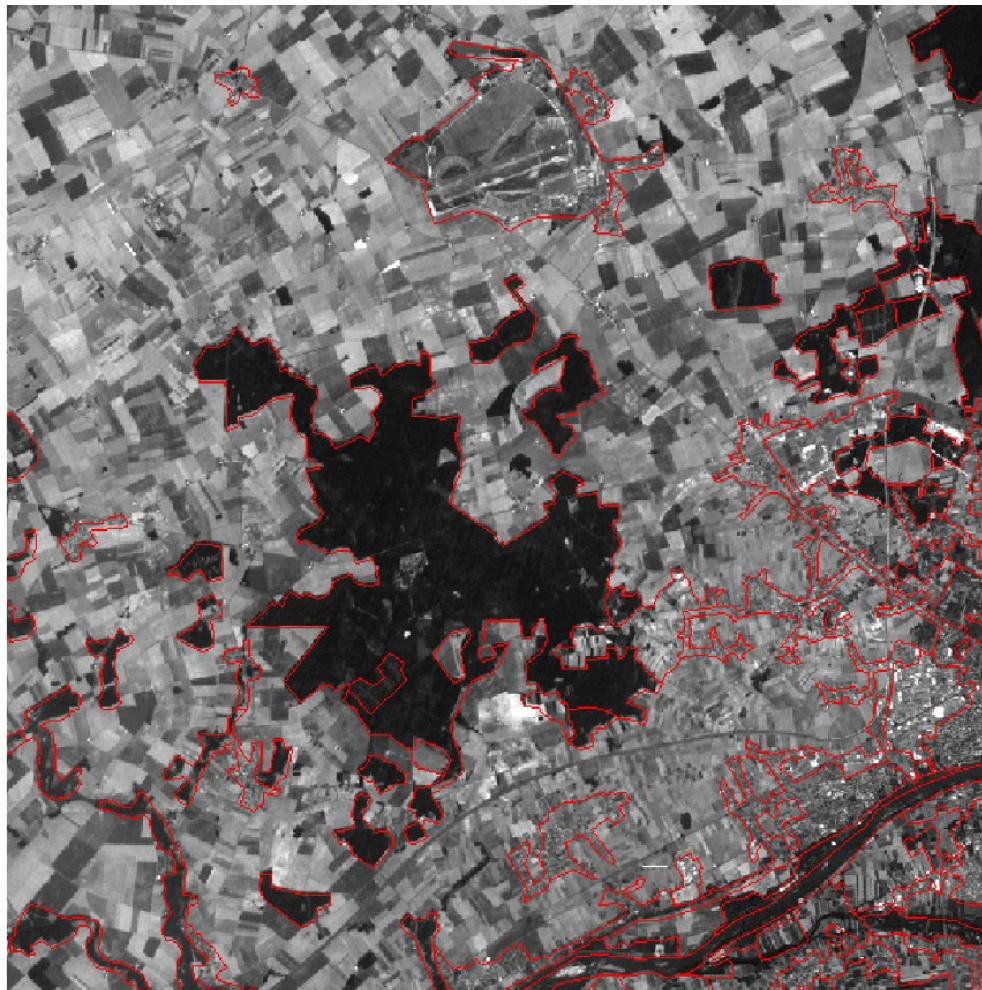


FIG. 4.3 – Couche CLC vectorielle, superposée à l'image SPOT correspondante sur la région du Loiret (Bois de Bucy) extraite de la scène complète de la figure 4.2.

vectorielles, ceci pour juger de la correspondance entre les images et les cartes.

- produire une couche raster “masque” à partir des cartes CLC pour isoler une zone d’intérêt. Par la suite, la visualisation des images ne fait apparaître que la partie de l’image correspondant à cette zone.
- manipuler, analyser et produire des couches de données raster par des opérations arithmétiques et logiques.
- développer des scripts à partir des modules de GRASS, et en particulier pour l’automatisation du processus de génération des zones d’intérêt et des images satellites correspondant à ces zones, et l’exportation des résultats (format raster) obtenus, pour une utilisation éventuelle en Matlab ou en C.

La figure 4.2 montre la couche CLC vectorielle du Loiret, superposée aux deux scènes SPOT couvrant des parties de la même région, tandis que la figure 4.3 montre une sous-partie de ce recouvrement.

TAB. 4.2 – Les 22 classes CLC présentes dans la région de test ainsi que le nombre de pixels pour chaque classe.

CLC	Nomenclature	Nombre de pixels
111	Tissu urbain continu	4231
112	Tissu urbain discontinu	161008
121	Zones industrielles et commerciales	37849
122	Réseaux routiers et ferroviaires et espaces associés	2878
124	Aéroports	17412
131	Extraction de matériaux	2283
141	Espaces verts urbains	4014
142	Equipements sportifs et de loisirs	4919
211	Terres arables hors périmètres d'irrigation	1105189
221	Vignobles	687
222	Vergers et petits fruits	17972
231	Prairies	3063
242	Systèmes culturaux et parcellaires complexes	93528
243	Surfaces essentiellement agricoles interrompues par des espaces naturels importants	16018
311	Forêts de feuillus	227825
312	Forêts de conifères	23648
313	Forêts mélangées	12062
324	Forêt et végétation arbustive en mutation	6309
331	Plages, dunes et sable	2042
411	Marais intérieurs	1165
511	Cours et voies d'eau	14236
512	Plans d'eau	2198

4.1.4 Description de la zone de test

Nous avons choisi de travailler sur la zone du Loiret qui est géoréférencée. En fait, la zone-test est une partie de la scène 39-253 (scène à gauche sur la figure 4.2) : elle est délimitée en bleu sur la figure 4.2.

La région test choisie contient 22 classes CLC, réparties comme indiqué dans le tableau 4.2. Elle a une taille de 1226×1436 pixels, c'est-à-dire de 1 760 536 pixels. Rappelons que la taille d'un pixel est de $20m \times 20m$.

Nous observons que ces classes sont très inégalement réparties dans l'image. Ceci est mis en évidence dans la figure 4.4, qui montre les populations des classes les plus nombreuses triées par taille et sur la figure 4.5 le pourcentage de l'image traitée si l'on se restreint aux n classes les plus nombreuses.

La classe des *terres arables* est largement majoritaire dans la scène (plus de 62 % des surfaces). Suit la classe des *forêts de feuillus*, puis, par ordre décroissant, le *tissu urbain discontinu*, les *systèmes parcellaires complexes* et les *zones industrielles et commerciales*. Nous notons que la classe la moins peuplée, celle des *vignobles* dispose tout de même de 687 échantillons (suffisamment pour procéder à un apprentissage), même si sa représentativité n'excède pas 4 pour 10 000.

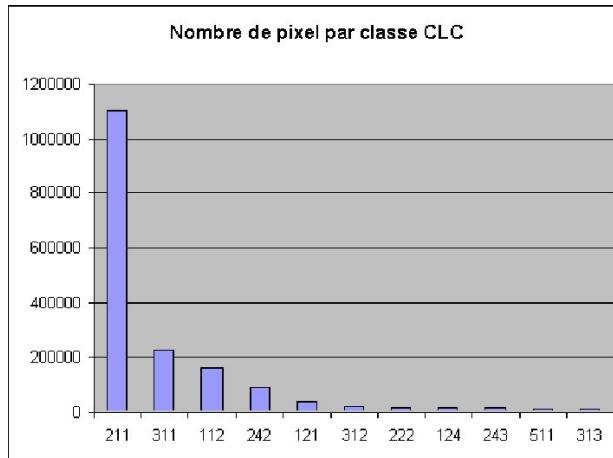


FIG. 4.4 – Les classes ordonnées par leur effectif. La classe des *terrains arables* est 5 fois plus représentée que la seconde classe, celle des *forêts de feuillus*.

Nous notons qu'un certain nombre de classes ne se distinguent que par une analyse contextuelle de l'image et ne pourront probablement pas être séparées par une classification sur la seule radiométrie. Il s'agit :

- des classes 511 et 512 : *cours et voies d'eau et plans d'eau*.
- de la classe 313 et l'ensemble {311 + 312} : *forêts mélangées* et de l'ensemble {*forêts de feuillus + forêts de conifères*},
- de la classe 242 (*systèmes culturaux et parcellaires complexes*) et la classe 243 (*surfaces agricoles interrompues*),
- des classes 112 et 121 : *tissu urbain discontinu et zones industrielles et commerciales*,

Enfin, des échantillons appartenant aux classes 124 (*aéroports*) et 142 (*équipements sportifs*) pourront basculer dans des classes différentes : 231 (*prairies*) ou 122 (*réseaux routiers ou ferroviaires*), voire 121 (*zones industrielles et commerciales*) par une classification non-contextuelle.

4.2 Méthodologie d'apprentissage et de classification

4.2.1 Protocole

La méthodologie que nous nous proposons de suivre est la suivante :

1. CLC nous propose une partition de l'image et une classification de chaque région. Nous allons accepter tout d'abord cette partition et cette classification et déterminer les éléments qui nous permettent de retrouver cette classification à partir de l'image seule, sur de simples critères de radiométrie. Nous aurons ainsi à déterminer :
 - les primitives de l'image qui sont les plus discriminantes,
 - les méthodes de classification les plus pertinentes.
2. Nous serons probablement alors amenés à remettre en cause certaines classes qui sont inaccessibles avec nos outils (éliminer des classes trop rares, fusionner des classes trop semblables, répartir les classes correspondants aux mélanges). Nous nous appuierons pour cela sur les remarques faites à la section 4.1.4, ainsi que sur

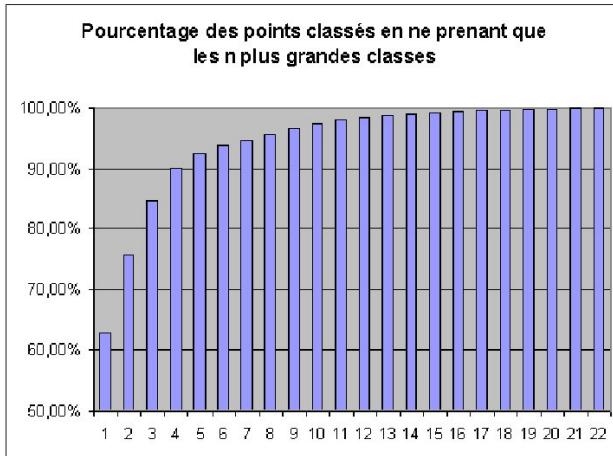


FIG. 4.5 – Le pourcentage de la surface de l'image couverte par les n classes les plus riches de la scène.

les confusions créées sur la zone test à l'étape 1 ci-dessus. Après ces simplifications, nous reprendrons l'apprentissage 1 sur un corpus de classes plus réduit.

3. Nous remettrons alors en cause la partition en nous appuyant sur les spécificités de CLC (simplification des formes, mélanges assimilés à des classes pures, etc.). Ceci nous amènera probablement à des déterminations plus précises des classes pures.
4. Enfin nous nous attacherons aux classes complexes qui ne se discriminent que par des primitives plus compliquées (par exemple des primitives de texture) ou par une analyse macroscopique de la classification mettant en oeuvre des éléments de géométrie (par exemple la classe des *champs mélangés*).

Dans toute cette démarche, nous serons amenés à définir de nombreux éléments de la méthodologie : les primitives issues de l'image (R, V, PIR, IB, NDVI, ISU, voire primitives de texture), la méthode de classification, les modèles de données (explicites ou paramétriques), les critères de qualité de la classification (taux global de fausse classification, trace de la matrice de confusion, coût bayésien, erreur moyenne, erreur maximale, ...), la stratégie globale de reconnaissance (arbre de décision, partition, un contre tous, ...).

Bien sûr, une démarche optimale devrait déterminer tous ces éléments simultanément, mais la complexité de cette tâche nous semble nécessiter une démarche séquentielle faite de fréquents retours aux optimisations précédentes afin d'évoluer vers un "optimum raisonnable".

4.2.2 Hypothèses

Nous nous plaçons tout d'abord dans une hypothèse de stationnarité des processus aléatoires, c'est à dire qu'une même classe C_i ($i = 1, \dots, N$), représentée dans l'image par n_i régions R_j^i ($j = 1, \dots, n_i$) se caractérise par sa probabilité à priori p_i , constante dans toute l'image et aura pour chaque région la même probabilité conditionnelle de classe $p(x|i)$, où x est toute mesure faite en un pixel d'image (les intensités R, V ou PIR, les mesures ISU, NDVI ou IB, etc.). Nous faisons également l'hypothèse que les pixels d'une même région sont indépendants les uns des autres (de même, a fortiori, que les pixels

adjacents de 2 classes différentes). Nous constaterons certainement les limites de cette hypothèse qui conduit traditionnellement à des effets “poivre et sel” dans les classifications. Nous y pallierons alors par des méthodes contextuelles.

4.2.3 Combinatoire des choix

Les primitives Comme précisé précédemment, nous choisissons de n'examiner tout d'abord que les mesures ponctuelles radiométriques faites sur un seul pixel, à l'exclusion des primitives de texture ou de structure.

Nous disposons du vecteur multispectral {PIR, R, V}, et calculons de plus, l'indice de végétation NDVI, l'indice de brillance IB ainsi que l'indice de bâti ISU, comme décrits dans la section 3.2.1.

Il est à noter que les indices sus-présentés devraient être calculés à partir des réflectances issues des différents canaux et non à partir des radiométries calibrées. Ce ne sont donc que des “pseudo-indices”, car ne disposant pas de ces données, nous avons utilisé les niveaux de gris.

Le nombre d'attributs pour chaque pixel n'étant pas très élevé (6 attributs pour chaque pixel), nous pouvons nous permettre de faire des combinaisons manuelles, afin d'identifier l'ensemble de primitives le plus discriminant. Nous établissons donc 3 ensembles de primitives qui nous permettront de juger de l'influence des indices de combinaison des bandes spectrales :

- PIR, R et V noté {EP1}
- PIR, R, V, NDVI et IB noté {EP2}
- PIR, R, V, NDVI, IB et ISU noté {EP3}.

Les modèles de distribution des radiométries Sous l'hypothèse de stationnarité des distributions, nous pouvons adopter 2 approches pour représenter les mesures caractérisant une classe :

- un modèle explicite dans lequel une primitive x prenant les valeurs $r \in \mathbb{R}$ se caractérise par une loi de probabilité conditionnelle à la classe i représentée par les valeurs $p_i(x = r), \forall r \in \mathbb{R}$.
- un modèle paramétrique qui sera en l'occurrence gaussien et se réduira donc au doublet \bar{x}, σ_x moyenne et variance de la classe.

Nous serons cependant amenés à travailler avec des variables de dimension 3 à 6, qui conduiraient, dans le modèle explicite à des représentations très coûteuses en stockage et très peu fiables en estimation. Nous choisirons donc systématiquement la représentation paramétrique gaussienne. Nous testerons la gaussianité des variables aux diverses étapes du traitement. En effet, nous pouvons nous attendre à rencontrer des classes très inhomogènes lors des premières classifications, en raison du principe même de construction de CLC. Nous améliorerons alors les estimations en rejetant les éléments les moins probables sur divers critères et itérerons le processus.

La méthode de classification Comme indiqué au début de ce chapitre, la technique utilisée pour l'apprentissage et la classification repose sur la classification bayésienne. Ici, elle consiste à maximiser la probabilité à postériori de chaque mesure faite en un pixel de l'image. Cela revient à maximiser $p(x|C_i)p(C_i)$, le modèle étant supposé gaussien (voir équation 3.24). Lorsque la distribution a priori est uniforme, la classification est basée sur le maximum de vraisemblance.

Evaluation de la qualité Dans ce chapitre, la qualité de la classification est évaluée de plusieurs façons :

1. Des critères subjectifs par examen visuel des cartes de classification automatique.
2. un critère objectif complet : la matrice de confusion $M(C_i, C_j)$ qui exprime la probabilité de reconnaître dans la classe C_j , un pixel issu de la classe C_i . Mais cette matrice très volumineuse est difficile à exploiter (plus de 400 termes pour les 22 classes initiales de la zone test). On peut en extraire des mesures plus synthétiques ;
3. la trace T_M de $M(C_i, C_j)$ divisée par N le nombre de classes, que l'on souhaite proche de 1 :

$$T_M = \frac{\sum M(C_i, C_i)}{N}$$

4. le même critère pondéré par la population des classes, \dot{T}_M :

$$\dot{T}_M = \frac{\sum P(C_i)M(C_i, C_i)}{\sum P(C_i)}$$

5. l'entropie jointe entre la classification CLC, C , et la classification finale apprise, C' . Il s'agit de la variation de l'information [Meila, 2002] :

$$VI = H(C) + H(C') - 2I(C, C')$$

avec $H(C)$ et $H(C')$, les entropies des classifications C et C' , et $I(C, C')$, l'information mutuelle entre elles.

$H(C) = -\sum p(C_i) \log p(C_i)$, et $I(C, C') = \sum p(C_i, C'_i) \log p(C_i|C'_i)$, où C_i est la classe i dans CLC et C'_i la classe i dans l'image.

Prise en compte de l'information contextuelle La définition de l'homogénéité d'une classe selon CLC n'étant pas absolue (au moins 75% de la superficie de l'unité paysagère doit appartenir à une même classe d'occupation du sol)[Gregorio & Jansen, 1998], la plupart des classes CORINE Land Cover sont très mélangées dans l'espace des caractéristiques. Nous avons donc mis en oeuvre diverses manières d'éliminer les pixels ayant les plus faibles probabilités d'appartenance à une classe, ceci afin d'améliorer les estimations.

- un vote majoritaire dans une fenêtre 3×3 , qui procède par un balayage arbitraire de l'image et attribue un pixel à la classe la plus représentée ;
- un procédé itératif qui consiste à recalculer la moyenne et la matrice de covariance, après élimination des pixels ayant une probabilité d'appartenance à la classe inférieure à une probabilité seuil. La probabilité seuil π est calculée comme :

$$\pi = p(x = \bar{x} \pm 3\sigma_x)$$

permettant ainsi de garder plus de 99% de l'information. Les calculs sont répétés jusqu'à stabilisation de la moyenne.

- les champs markoviens, qui bénéficieront de la connaissance d'un modèle d'attaché aux données (ici gaussien) et d'une connaissance des paramètres du terme d'attaché aux données (la moyenne m_j et la variance σ_j de la classe j). L'approche markovienne utilisée ici est basée sur l'algorithme des modes conditionnels itérés (*Iterated Conditional Mode* ou ICM) décrit dans la section 3.7.

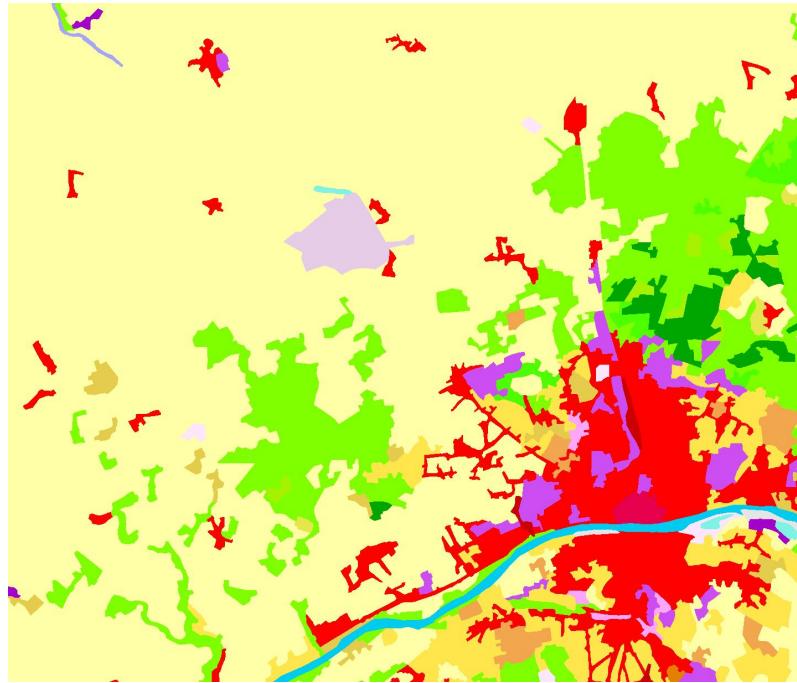


FIG. 4.6 – *Région 1* : carte CORINE Land Cover représentant la région de test contenant 22 classes.

Nous avons considéré les cliques de second ordre associés à un 8-voisinage et utilisé un modèle de Potts [Wu, 1982] pour calculer la fonction d'énergie $U(x_s)$, x_s étant l'étiquette en un site (pixel) s :

$$U(x_s) = -\ln(P(y_s/x_s)) + \sum_{t \in V_s} \beta \delta(x_s, x_t) \quad (4.1)$$

où $\delta(x_s, x_t)$ est le symbole de Kronecker, β est un terme de régularisation, et V_s est le système de voisinage pour le site s .

4.3 Classification sur critères radiométriques

Les tests ont été menés avec la classification bayésienne pour les 22 classes CLC présentes dans la région du Loiret de la figure 4.6, et sur les 3 ensembles de primitives : {EP1}, {EP2} et {EP3}. Ceci permettra d'avoir une idée des attributs spectraux discriminant le mieux une classe donnée. La prise en compte du voisinage d'un pixel lors de la classification est effectuée via les champs markoviens, en utilisant l'algorithme ICM tel que décrit précédemment.

Les tests de l'apprentissage fait par une approche bayésienne avec le Maximum a Posteriori (MAP) donnent les résultats exposés dans les tableaux 4.3, 4.4 et 4.5 respectivement pour les ensembles d'attributs {EP1}, {EP2} et {EP3}. Un exemple visuel de résultat avec {EP1} est présenté dans la figure 4.7.

TAB. 4.3 – Matrice de confusion exprimée en pourcentage, obtenue avec le Maximum a Posteriori et les primitives {EP1} pour les 22 classes. Les lignes représentent les classes réelles et les colonnes, les classes estimées.

clé	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512	
111	44	16	4		2				4		2				4					26			
112	3	42	2			1			31		2				1	11				1			
121	5	36	8			6			25		2				3	2	7			3			
122	23	37	2			1			32						1	2				1			
124	2	12	3			1	3		56						5	1	6	11		1			
131	8	7	1			29			38						9	2	2	2	1	1			
141	1	16				1			13		4	2			7	43	8			3			
142	2	10	1			2			29		4	3			1	18	17	6		4	2		
211	1	7				1	2		78		1		2		2	2	4			2			
221		5							37		1	11			6	27		13			2		
222		5							35		4	3			7	31		11			1		
231		5							50		25				7	8	1	3					
242		14				2			50		1	4			1	3	15	9			1		
243		11				2			33		3	3			3	28		12			1	2	
311	2						4		3	1					75	10	1	2		1			
312	1						2		3						60	5	28						
313	2						6		2	3					71	6	7	1					
324							7		4	8					60	11	8	1		1			
331	4	14	2			1	5		31		1	1			14	1	17	3	1	7			
411	4	8	1						20		9	1			1	19	11	2			37		
511		5							5						1	1	8	1			69		
512	67	10							12						2					7			

TAB. 4.4 – Idem avec le Maximum a Posteriori et l'ensemble de primitives {EP2} pour les 22 classes.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	37	14	7			1			2			9			1		3			34	1	
112	3	38	4			1			14			7			1		12			2		
121	6	29	18			1	2		11			4			3		7	1	9	1	1	
122	30	30	8	6		1			13			2			1		2					
124	2	11	4		3	1			13			23			2		10		29			
131	9	7	2		1	26			18			7			20		2		5	1		
141	14					1			3			1	4		1		16	41	1	12		
142	2	9	1		1	1			11			1	8		4		24	17	1	12		
211	1	7	1		1		2		58			1	6		5		1	2	9	4		
221	4								4			22			8		36	1	26			
222	4								11			2	12		1		14		33	1		
231	5								25			34			7		12	2	13		3	
242	12	1			1	1			22			1	17		2		5	18	19	1	1	
243	9				1	1			13			1	10		2		8	30	21	2	2	
311	2								1			1	1		78		9	4	2	1		
312	1											3			39		7	50	1			
313	2											2			64		8	18	2			
324												1	4		58		13	22	1	1		
331	3	13	4		1	2			13			7			20		2	16	6	1	8	
411									9			2	5		33		11	1	4	36		
511	3	7	1		1				2			1	3		2		7	2		70	1	
512	66	5	4		1	1			7			1	2		2		2			7		

TAB. 4.5 – Idem avec le Maximum a Postériori et l'ensemble de primitives {EP3} pour les 22 classes.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	69		9	1	3				6	2	2				1					2	2	
112	4		4	2	1				27	18	5	1	20		17					1	1	
121	6		8	3	5				27	12	12	2	10	1	11					1	1	
122	29		21	1	1				27	6	3	5	6									
124	2		1	1	1				10	32	8	11	32									
131	11		1	1	21				9	15	33	4	6									
141	3		1						7	1	8	1	14	47	5	10			1	1		
142	4								6	2	15	9	22	19	4	13			4	1	1	
211	1		1						36	3	14	24	1	3	11					5		
221									2	25	1	6	45	3	18							
222	1								5	4	19	2	13	34	3	17				3		
231									18	41	1	6	15	4	14							
242	1		1						15	1	27	6	4	22	2	19			1			
243	1								9	2	16	6	7	34	1	20			2	1		
311									1	1	2	1	70	10	12	2			1			
312										1	3	19	7	70	1							
313									2		3	1	41	8	42	2						
324										2	4		47	11	34	1			1			
331	9		3	1	1				10	1	15	28	2	18	1	8			2	1		
411									5	4	7	2	31	10	2	4			35		46	1
511	26		2						3	2	5	2	8	1	2							
512	71		4						7	4	5	3	4	4	3							

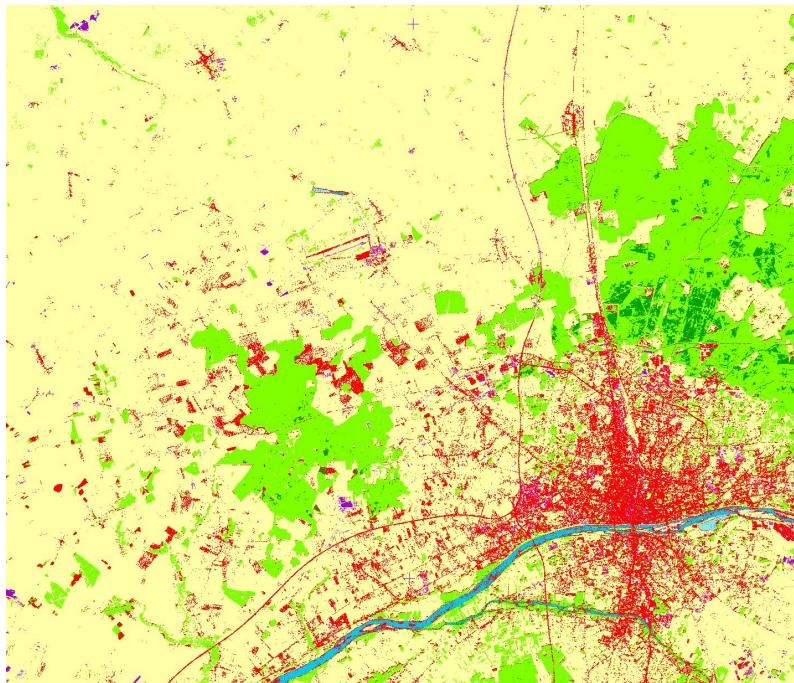


FIG. 4.7 – Résultat de la classification obtenue avec le Maximum a posteriori et les primitives {EP1}. Le bruit dit “poivre-sel” présent dans l’image classifiée, est dû au traitement suivant une approche pixel.

4.3.1 Analyse des résultats

En examinant ces matrices, nous constatons que la diagonale de la matrice est généralement très faible et il existe plusieurs classes qui ne sont quasiment pas identifiées (124, 141, 142, 222, 242, ...). Ces mauvaises detections peuvent s’expliquer par plusieurs causes :

Classes rares : Certaines classes sont très peu représentées. Une classification bayésienne aura tendance à mal les classer chaque fois qu’elles seront en conflit avec des classes plus peuplées. C’est probablement le cas des classes 124 (*aéroports*), 141 (*espaces verts urbains*), 142 (*équipements sportifs*) ci-dessus.

Mélanges : Cela était prévisible pour certaines classes comme celles dites “à caractère mixte” (242, 243) [Ins, 2005b] qui sont des regroupements de catégories d’occupation des terres, dont la superficie pour chaque catégorie ne satisfait pas le critère de superficie minimale des unités cartographiées (25 ha), pour être considérée comme une classe CLC à part entière. Et donc, les pixels que nous classifions sont naturellement catalogués dans la classe CLC correspondante, et non dans celle de regroupement. Ainsi, les pixels de la classe 242 (*Réseau de cultures variées*) sont classifiées dans les classes 312, 211 et 324. Il faudra donc certainement tenir compte des critères spatiaux et contextuels pour classifier plus correctement ce type de classes.

Classes texturées : D’autre part, étant donné que nous avons utilisé uniquement des attributs spectraux, des classes non ou peu identifiées peuvent provenir de la confusion

avec des classes ressemblantes radiométriquement, mais distinguable sur leur texture. C'est probablement le cas de la végétation : les classes 221 (*vignobles*) et 222 (*vergers*) ne peuvent se distinguer des *forêts* ou *prairies*.

Classes indiscernables : Il est aussi à noter que les images satellitaires ne fournissant pas d'information pertinente sur la hauteur des arbres, les classes 3.1 (*forêts*) et 3.2 (*arbustes, broussailles et savanes*) sont donc souvent confondues. De plus la classe 324 (*forêt en mutation*) par exemple, est naturellement classifiée dans 311, 312 ou 313, les 3 classes de forêts *feuillus*, de *conifères* et *mélangées* les plus représentées.

Malgré ces défauts prévisibles et attendus, nous notons qu'il existe quelques classes assez bien reconnues telles 111 (*tissu urbain continu*), 311 (*forêts de feuillus*), 211 (*terres arables*) et 511 (*cours et voies d'eau*). Ces zones représentant plus de 75 % de la surface de l'image, nous pouvons espérer une classification raisonnable (mais grossière).

Certains indices décrivent mieux certaines classes. En particulier, le meilleur taux de reconnaissance de la classe 211 (*terres arables*) (78%) est obtenu avec l'ensemble de caractéristiques {EP1}, tandis que {EP2} est le plus pertinent pour reconnaître les *fôrets de feuillus* (classe 311), et avec {EP3}, on passe de 44% à 69% pour le *tissu urbain continu* (classe 111).

4.3.2 Comparaison de classifications

Par ailleurs, les différentes classifications que nous avons obtenues ont été comparées à la carte CLC, à l'aide de VI et \dot{T}_M . La variation de l'information est une métrique : plus elle est petite, plus la classification est proche de celle de CORINE Land Cover ; tandis que \dot{T}_M tend vers 1 lorsque la classification se rapproche de celle de CORINE Land Cover.

Nous avons obtenu les résultats du tableau 4.6 pour ces tests. Nous remarquons que d'après VI et \dot{T}_M , la classification conduite avec {EP1} est la plus proche de CLC.

TAB. 4.6 – Critères de comparaison de classifications obtenues avec le MAP, pour chaque ensemble de caractéristiques.

	Critères	{EP1}	{EP2}	{EP3}
MAP	VI	2.11	2.52	2.66
	\dot{T}_M	0.63	0.52	0.33

Observant visuellement les classifications, on constate l'importance des inclusions de très petites zones dans des classes. La nécessité de mettre en œuvre une procédure contextuelle semble donc s'imposer.

4.3.3 Traitement contextuel

Le traitement post-classification que nous avons ensuite appliqué à l'image classifiée avait pour but de tenir compte du voisinage des pixels pour éliminer l'aspect "poivre-sel" (homogénéiser les régions) de l'image classifiée, comme le montre la figure 4.8 obtenue avec les caractéristiques {EP1}. Par ailleurs, la matrice de confusion exprimée en pourcentage, pour le même ensemble d'attributs est consignée dans le tableau 4.7. Celles des

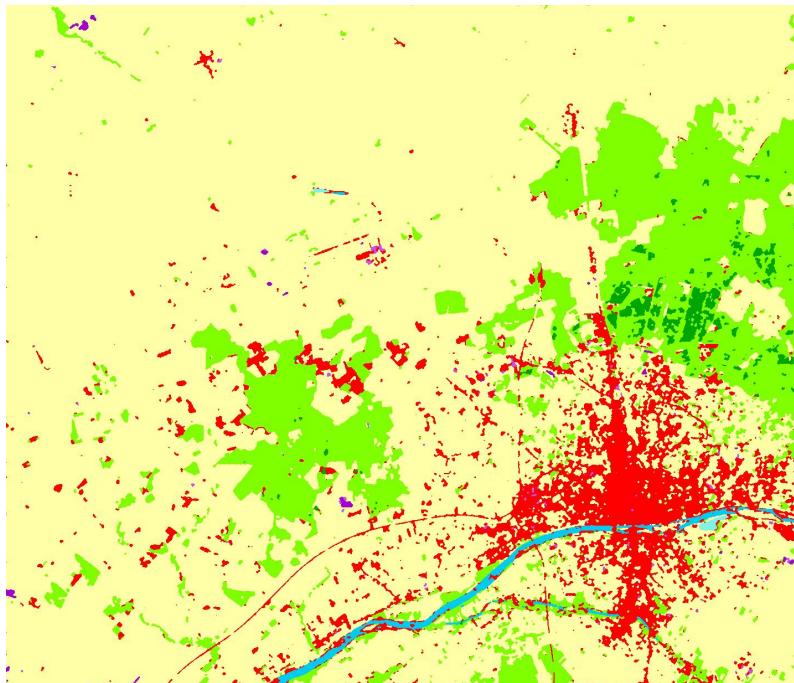


FIG. 4.8 – Résultat de la classification obtenue avec les primitives {EP1} et le Maximum a postériori (figure 4.7), suivie d'un traitement contextuel par champs de Markov.

ensembles {EP2} et {EP3} sont données en annexe A.3, dans les tableaux A.1 et A.2 respectivement.

Nous remarquons que, par rapport au cas précédent, le taux de reconnaissance des classes qui étaient déjà assez bien identifiées augmente de manière notable. Par exemple pour la classe 111, on passe de 69% à 91% avec {EP3}. Cependant, concernant les fortes confusions entre les classes, nous remarquons que plusieurs pixels d'autres classes sont catalogués comme appartenant à 211, 311, 312 ou encore 112. Ceci pourrait être dû au fait que ces dernières classes sont bien représentées (nombre élevé de pixels) et donc, ont tendance à absorber les petites classes.

Tout comme dans le cas précédent, nous observons que certains indices mettaient mieux en évidence certaines classes. Le tableau 4.8, donne les classes reconnues à plus de 80%, à plus de 40% et pas reconnues du tout, pour chaque ensemble de caractéristiques.

En d'autres termes, pour isoler la classe 211 par exemple (*Terres arables hors périmètre d'irrigation*) et espérer avoir un taux de reconnaissance de plus de 80%, nous pouvons utiliser une classification MAP+voisinage et les caractéristiques {EP1}. De même, avec le même classifieur et les caractéristiques {EP3}, nous ne pourrons jamais isoler la classe 112 ou la classe 231.

Les indices de qualité, calculés pour comparer ces classifications avec CLC donnent les résultats du tableau 4.9.

Il en résulte que d'après VI et \dot{T}_M , les classifications bayésiennes avec le MAP suivi de la prise en compte du voisinage des pixels (MAP+voisinage) sont plus proches de la classification de CLC que celles obtenues avec le MAP seul. En outre, VI et \dot{T}_M indiquent qu'avec le MAP+voisinage, la classification la plus proche de CLC est réalisée avec l'en-

TAB. 4.7 – Matrice de confusion exprimée en pourcentage, obtenue avec le MAP+voisinage et les primitives {EP1} pour l'élimination de l'aspect "poivre et sel".

TAB. 4.8 – Classes reconnues à plus de 80%, à plus de 40% et à 0%, pour chaque ensemble de caractéristiques, en utilisant le classificateur MAP+voisinage.

	{EP1}	{EP2}	{EP3}
+ de 80 %	211	311, 511	111
+ de 40 %	112, 411		231, 313
0 %	122, 141, 221, 242, 512	141, 242	112, 121, 124, 141, 142, 221, 231, 242

TAB. 4.9 – Critères de comparaison de classifications calculés avec le MAP+ voisinage pour l'élimination de l'aspect "poivre et sel", et chaque ensemble de caractéristiques.

	Critères	{EP1}	{EP2}	{EP3}
MAP +voisinage	VI \hat{T}_M	1.76 0.69	2.17 0.59	2.47 0.37

semble d'attributs {EP1}.

4.3.4 Remarques

- la classe 111 (*tissu urbain continu*) est souvent classifiée en 511 (*cours et voies d'eau*).
- le *tissu urbain discontinu* (112) se retrouve souvent en 211 (*terres arables*, ce qui est logique si l'emprise des villages est large).
- les classes 121 (*zones industrielles*) et 122 (*réseaux routiers*) sont reconnues en 111 (*tissu urbain continu*) et 112 (*tissu urbain discontinu*).
- la classe *aéroports* (124) est classifiée en *terres arables* et *prairies*.
- la classe 211, la plus peuplée, absorbe toutes les classes de terres cultivées et de prairies, ce qui semble logique,
- une bonne partie des classes 221 (*vignobles*), 222 (*vergers et petits fruits*) et 243 (*surfaces agricoles interrompues par des espaces naturels*) sont classifiées dans 312 (*fôrets de conifères*) et 324 (*fôrets et végétation arbustive en mutation*). Tout comme la classe des *fôrets mélangées* (313), ces dernières sont à leur tour absorbées par la classe 311, la plus peuplée des fôrets. L'autre partie des classes 221, 222 et 243 est cataloguée comme étant des *terres arables*.
- la classe 512 (*plans d'eau*) est systématiquement mal classée en 111 (*tissu urbain continu*) (on peut s'interroger sur la dimension de ces étendues d'eau).
- Les *marais intérieurs* (411) sont partiellement classés en 311 (*fôrets de feuillus*) et 211 (*terres arables*).

Il est donc assez clair que la classification vérifie bien les défauts que nous attendions dès l'analyse de la cartographie locale, en section 1 :

1. les zones mélangées sont distribuées dans leurs composantes élémentaires,
2. la classification bayésienne, en cas d'ambiguïté, favorise les classes les plus peu-peuplées,
3. les zones texturées sont mal distinguées avec nos seuls critères radiométriques.

Des conclusions supplémentaires apparaissent :

1. les zones de cours d'eau sont confondues avec les zones urbaines,
2. les zones de marais n'ont pas de signature radiométrique discriminante.

4.3.5 Solutions possibles

Regroupement de classes Une première solution consiste à s'appuyer sur les remarques ci-dessus et regrouper les classes comme présenté dans le tableau 4.10.

TAB. 4.10 – Regroupement des classes en fonction des mélanges.

Classes CLC	Groupe
111 + 512	{1} : urbain continu
211+231+242+124	{2} : champs
311+ 312+313+221 +222+324+243	{3} : forêts
112+121+122	{4} : urbain discontinu
511	{5} : eau
131+141+142+331+411	{6} : autres

Ainsi, si l'on procède à ces regroupements, on obtient le tableau de confusion présenté en 4.11, bâti à partir de la table 4.7 en réunissant les lignes et les colonnes conformément aux conventions de la table 4.10.

TAB. 4.11 – Construction de classes artificielles par regroupement de classes CLC conformément au tableau 4.10 ; classification par MAP, avec le jeu de primitives {EP1} et élimination de l'effet "poivre et sel".

	{1}	{2}	{3}	{4}	{5}	{6}
superclasse {1}	68,50	6,50	4,50	5,00	15,50	0,50
superclasse {2}	0,50	66,50	26,83	3,83	0,33	1,33
superclasse {3}		20,33	77,83	0,83	0,33	0,33
superclasse {4}	30,00	19,67	6,67	32,67	7,67	2,00
superclasse {5}	1,00	4,00	11,00	3,00	78,00	2,00
superclasse {6}	15,33	28,17	32,17	3,67	4,33	14,00

On voit que 4 superclasses sont assez bien reconnues : l'urbain continu, les champs, les forêts et l'eau. Comme il se doit l'urbain discontinu est également dispersé entre urbain continu, champs et urbain discontinu (ce qui prouve la nécessité de déterminer cette classe sur des critères plus macroscopiques). Enfin la classe "autres" reste dispersée.

Méthodes hiérarchiques Une autre solution est de procéder de façon hiérarchique, par exemple en éliminant les classes trop rares et en isolant celles qui sont "correctement" identifiées. Cela pourrait aboutir à une classification par arbre de décision, c'est à dire

que nous pourrions construire un arbre de manière à identifier chaque classe CLC l'une après l'autre, ceci en utilisant bien sûr d'autres attributs et/ou d'autres classifieurs.

Nous approfondirons ces directions de recherche dans la section 4.5.

Par ailleurs, l'inégale répartition de ces classes a entraîné un phénomène d'absorption des classes très peu représentées par les plus peuplées (favorisé par la classification bayésienne). De plus, l'utilisation des seules caractéristiques radiométriques s'est avérée insuffisante pour l'identification de certaines classes riches en texture et ainsi pour éviter les confusions entre certaines catégories. Une prise en compte des paramètres de texture s'impose donc pour améliorer cette classification.

4.4 Prise en compte des textures

4.4.1 Caractéristiques QMF

Parmi les grandes catégories d'attributs de texture, nous avons utilisé les caractéristiques QMF ou *Quadratic Mirror Filters* [Mallat, 2003], décrites dans la section 3.2.2. De précédents travaux du Centre de Compétences ont montré que ces primitives basées sur les filtres en quadrature de phase donnaient de bonnes performances lors de l'indexation d'images satellitaires [Campedel et al., December 2004].

Pour extraire les paramètres QMF, la bande PIR a été découpée en de petites fenêtres. Travailant avec des images SPOT à 20 m de résolution, il nous a fallu mener quelques expérimentations pour ces images, alors que nos expériences antérieures étaient conduites en SPOT 5 Panchromatique à 5m. Nous avons essayé plusieurs tailles : 16×16 avec un pas de 8 pixels, 32×32 avec un pas de 5 pixels,... Expérimentalement, les imagettes de taille 32×32 avec un pas de 8 pixels ont donné les meilleurs résultats (qmf328). Dans l'expérimentation nous utilisons 2 échelles, ce qui donne 14 caractéristiques pour la moyenne et la variance de l'imagette, dont 8 sont utilisées (variances des coefficients des sous-bandes HH, HL, LH et LL, et moyenne de LL). Une interpolation bilinéaire est appliquée aux coefficients QMF pour revenir à la taille initiale de l'image.

Le tableau 4.12 donne les résultats de classification MAP sur les paramètres de texture uniquement, ainsi que la trace pondérée par la population des classes \bar{T}_M .

Nous constatons que certaines classes sont mieux reconnues, notamment celles du type 1 d'occupation du territoire (*Territoires artificialisés*). La classe 111 par exemple (*Tissu urbain continu*) est reconnue à 97%, avec très peu de confusions avec les autres classes. Il en est de même pour les *Vignobles* (classe 221).

Cependant, certaines classes restent dominées par 211 (*Terres arables*), 121 (*zones industrielles et commerciales*) et 324 (*Fôret et végétation arbustive en mutation*). De plus, les classes de forêts (311, 312, et 313) sont pratiquement inexistantes. En outre, la trace pondérée de la matrice de confusion (0.47) est faible par rapport au cas de la classification spectrale (0.69). Il pourrait être intéressant de combiner les caractéristiques pour bénéficier de la bonne identification des classes et d'une bonne classification globale.

4.4.2 Utilisation combinée des caractéristiques spectrales et texturelles

Les 8 coefficients QMF ont été concaténés à chaque ensemble de primitives spectrales ($\{\text{EP1}\}$, $\{\text{EP2}\}$ et $\{\text{EP3}\}$), et ce pour chaque pixel. Les résultats obtenus pour les ca-

TAB. 4.12 – *Région 1* : matrice de confusion avec le MAP et la texture seule (qmfb28) pour les 22 classes ; $\bar{T}_M=0.4690$. Les lignes représentent les classes réelles et les colonnes, les classes estimées.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	97	2																				
112	1	10	48	4	4	8	1	17	2						3					1		
121	3	66	3	7		4		13							1	2						
122	5	23	63			2		4	4													
124	6		61	1		24									4					2		
131	1		5	50	1	23									4					9		6
141	4	32	4	58	1	1		1							1					1		
142	11		12	1	31	21	1	1								5	1	14				
211	1	8	8		1	66	1	1	1						10	3						
221		2				1	97															
222	3	14	13		3	18	4		3						32	7						
231	3	18	3			50			3						20	2						
242	7	33	1	8	7	14	1	1	8						14	1	1	1				
243	3	31	5	4	3	23	1	1	2						25	2	1	1				
311		4	3	1	1	29	1		2						7	42						
312	2	27	7	1	1	26	1								34	2						
313		18	5	2		33	3		2						29	1						
324	1	7	2	11	3	7									83							
331															44	6	20					
411	3	6	11	8	3	26	1								11	2	88					
511															28	16						
512															1	83						

ractéristiques ($\{EP2\} + qmf328$) avec la prise en compte du voisinage à l'aide des champs de Markov sont consignés dans le tableau 4.13 pour ($\{EP2\} + qmf328$).

Ce tableau montre que certaines classes sont presque parfaitement identifiées : la classe 221 (*Vignobles*) à 99 % et la classe 111 (*Tissu urbain continu*) à 97%. Aussi, nous pouvons constater que, par rapport au cas de l'utilisation de la texture seule, les classes du type 1 d'occupation du territoire (*Territoires artificialisés*) sont mieux reconnues. C'est le cas de la classe 131 (*Extraction de matériaux*) dont le taux de reconnaissance passe de 50% à 85%.

Par ailleurs, les classes 324 et 211 sont bien reconnues, mais continuent d'absorber une partie non négligeable des pixels appartenant à certaines autres classes, entraînant ainsi des fausses alarmes. Ainsi la plupart des échantillons des classes de forêts (311, 312, 313) sont classifiés dans 324 (*Forêt et végétation arbustive en mutation*). Pareillement, les classes 242 (*Réseaux de cultures variées*) et 243 (*Surfaces essentiellement agricoles, interrompues par des espaces naturels importants*) sont réparties dans presque toutes les classes.

Ces confusions sont liées, soit à une similarité conjoncturelle et inopinée entre les classes, soit au fait que les classes font partie de celles dites "à caractère mixte", c'est-à-dire des classes mélangées par définition. Pour pallier ce problème, nous avons exploré et mis en œuvre plusieurs types de traitements qui sont vus dans la section suivante.

4.5 Amélioration de la classification

Parmi les méthodes déployées pour améliorer la classification et nous rapprocher sur certains points de l'approche conduite lors de la construction des cartes CORINE, nous avons testé, entre autres, la classification hiérarchique, le regroupement de classes et la représentation de l'image par graphes d'adjacence.

4.5.1 Classification hiérarchique

Le principe de cette classification, comme son nom l'indique, procède par des décisions successives qui isolent progressivement les classes difficiles à distinguer. Ce principe est schématisé dans la figure 4.9. Un premier test est effectué sur l'ensemble des échantillons. De la classification obtenue, les modèles des classes bien identifiées (suivant CORINE Land Cover) et ayant peu de fausses alarmes sont supprimés, ainsi que les pixels assignés à ces classes. Un autre test est appliqué sur les pixels restants et, de la même manière, les classes bien reconnues sont isolées. Cette opération est reconduite tant que des classes peuvent être "correctement" écartées. A chaque étape, il est possible d'utiliser un classificateur et des paramètres différents de ceux de l'étape précédente car, comme nous l'avons déjà vu, certaines caractéristiques décrivent mieux certaines classes et donc, nous pouvons espérer de meilleures performances de cette diversité. En outre, avant un test, des classes qui ont tendance à se mélanger peuvent être regroupées, ceci pour augmenter la capacité de reconnaissance. Cela pourra aboutir à une classification par arbres de décision, c'est-à-dire qu'une fois les petits groupes bien identifiés, et sachant quelles classes CLC sont incluses dans chaque groupe, il serait possible d'identifier chaque classe CLC l'une après l'autre, ceci en utilisant bien sûr d'autres attributs et/ou d'autres classifieurs.

Cependant pour pouvoir isoler une classe, il faut que cette dernière ait un bon taux de reconnaissance et peu de confusions avec les autres classes. Nous ne nous intéressons

TAB. 4.13 – *Région 1* : matrice de confusion avec le MAP + voisinage et l’ensemble de primitives ($\{\text{EP2}\}$ + qmf328) pour les 22 classes ;
 $\bar{T}_M=0.6413$

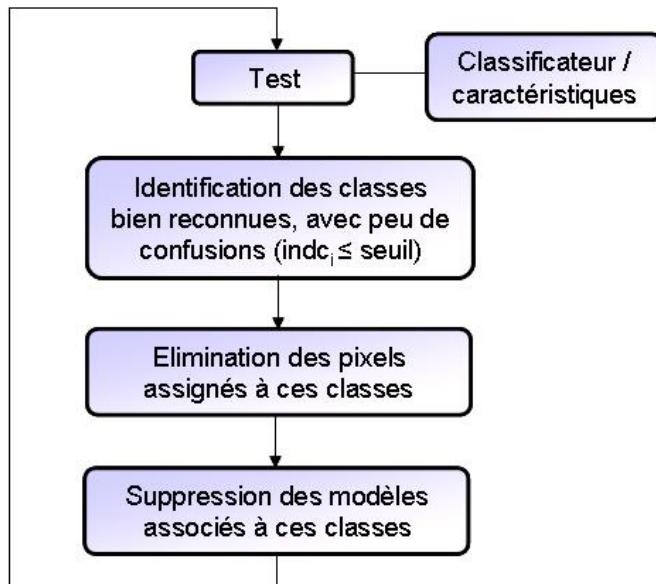


FIG. 4.9 – Principe de la classification hiérarchique : le test et seuil évoluent à chaque itération.

donc plus aux performances de la classification dans sa globalité, mais à celles de chaque classe individuellement. Outre le taux de bonne classification pour chaque classe (*producer's accuracy*), le critère utilisé pour la décision et défini pour chaque classe est le rapport de la somme des pourcentages des mauvaises classifications et des fausses alarmes sur le pourcentage de bonnes classifications. Ce critère ($indc_i$) est indépendant du nombre d'échantillons par classe et vaut 0 lorsqu'il n'y a pas de fausses alarmes, ni de mauvaises classifications.

$$indc_i = \frac{100 + \%n'_i - 2 \times \%n_{ic}}{\%n_{ic}} \quad (4.2)$$

où $\%n_{ic}$ est le pourcentage d'échantillons bien classifiés, et $\%n'_i$ est la somme des pourcentages des pixels assignés à cette classe.

Un exemple de résultats de tests effectués par cette méthode est donné par l'arborescence de la figure 4.10. Le seuil étant expérimentalement fixé à 0.35, toutes les classes dont le critère $indc_i$ est inférieur ou égal à 0.35 sont considérées comme isolables. L'arbre donne les pourcentages de bonne classification de chaque classe isolable, sous réserve de l'erreur du test précédent. Pour chaque niveau de la hiérarchie, nous avons utilisé soit {EP1}+qmf328, soit {EP2}+qmf328. Comme le montrent le taux de bonne classification et $indc_i$, un niveau de plus dans l'arbre permet d'isoler des classes qu'on ne pouvait bien identifier au niveau précédent. Des regroupements ont été effectués et les deux groupes du troisième niveau ont pu être dissociés. Cependant, dans chaque groupe, les classes restent emmêlées.

Une autre manière d'aborder cette classification hiérarchique est de considérer chaque classe mélangée (par exemple les classes 112 (*Tissu urbain discontinu*) ou 242 (*Réseau de cultures variées*)) et d'y effectuer un clustering. L'intérêt est d'isoler les différents constituants de ces classes hybrides et de trouver un modèle pour chaque constituant. Alors,

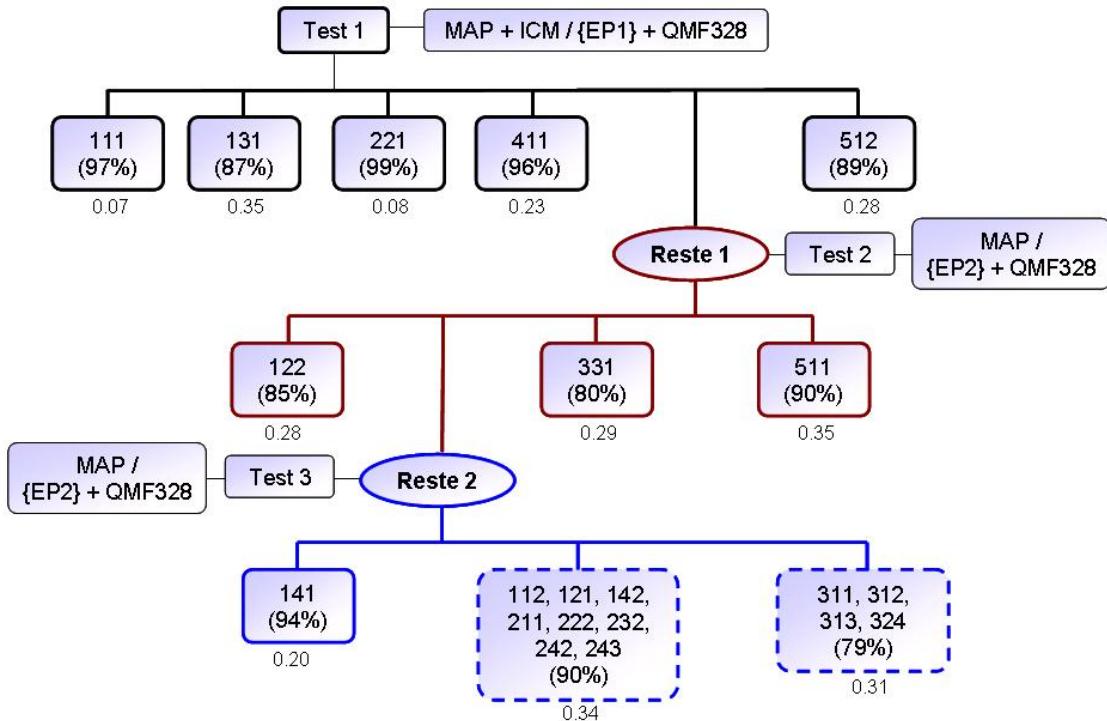


FIG. 4.10 – *Région 1* : résultats de la classification hiérarchique avec un seuil pour $indci$ fixé à 0.35. A chaque niveau de la hiérarchie, le pourcentage de bonne classification est donné pour chaque classe, sous réserve de l'erreur du niveau précédent. La valeur $indci$ est indiquée au-dessous de chaque bloc, plus elle est faible, meilleure est la décision.

lors de la classification, le modèle pour ce type de classes est remplacé par ceux des différents constituants. Cette procédure est une sorte de classification hiérarchique ascendante. Les tests effectués sur notre base de données ne nous ont pas donné entière satisfaction. En effet, les sous-classes obtenues par la méthode des k-means ne correspondent pas toujours aux différents constituants de la classe mélangée. Par exemple, la classe 112 (*Tissu urbain discontinu*) a été décomposée en 3 sous-classes, mais reste classée à 50% dans les *zones industrielles et commerciales*, tandis que les 3 petits groupes créés retiennent à peine 0%, 1% et 3% de bons échantillons. Ce problème peut être en partie résolu par la détermination automatique du nombre de clusters dans un ensemble de données, via des méthodes basées sur la théorie taux - distorsion [Sugar & James, 2003] ou de type AutoClass. Cependant, le risque qu'il y ait beaucoup de clusters est grand car aucune classe CLC n'est homogène à 100%.

4.5.2 Regroupement des classes

En s'appuyant sur les remarques faites après la classification dans la section précédente, les classes sont regroupées en hyperclasses. Au regard de la matrice de confusion, le regroupement des classes revient à inclure dans la classe dominante, les classes qui sont absorbées par cette dernière. Cependant, il arrive très souvent qu'une classe soit répartie de manière presque uniforme entre plusieurs classes (par exemple, les classes 242

et 243 dans le tableau 4.13). Alors, le choix pourrait se faire en considérant celle qui est la plus proche de la classe concernée suivant la classification hiérarchique de CORINE Land Cover. Cela entraînera certainement des confusions entre la nouvelle classe dominante et celles qui ont été écartées. A chaque fois, le modèle utilisé pour la classification est celui de la classe considérée comme dominante. Les modèles des autres classes constituant le groupe sont supprimés.

Les classes doivent être associées de manière “judicieuse” pour donner de bons résultats. Nous verrons s’il est possible de définir ce critère d’association des classes de manière automatique, au regard de la matrice de confusion.

Pour juger de la qualité de la classification et donc de ces regroupements, outre le critère $indc_i$ pour évaluer chaque classe séparément et la trace pondérée déjà utilisée, nous introduisons le paramètre $Kappa(\kappa)$, défini dans la section 3.5, qui est un indicateur global et plus objectif de la qualité de la classification.

Le tableau 4.14 et le tableau A.3 présenté en annexe A, exposent les résultats obtenus pour certaines combinaisons de regroupements. Les classes ont été associées en tenant compte des matrices de confusion précédemment obtenues et des similarités sémantiques entre les classes telles que définies par CLC. Le tableau 4.13 montrait qu'il existe de grandes confusions entre les postes des types d'occupation du sol 1 (*Territoires artificialisés*) et 2 (*Territoires agricoles*), et que globalement, les classes de type *Forêts* (type 3) sont mélangées entre elles. Un regroupement grossier basé sur ces remarques donne les résultats du tableau 4.14, avec deux hyperclasses ($H1=\{112, 121, 122, 124, 131, 141, 142, 211, 222, 231, 242, 243\}=211$, c'est-à-dire que nous avons gardé le modèle de la classe 211 pour cette hyperclasse et $H2=\{311, 312, 313, 324, 331\}=324$). Les tests effectués avec un regroupement un peu plus fin donnent les résultats exposés dans le tableau A.3.

TAB. 4.14 – *Région 1* : matrice de confusion obtenue en appliquant le MAP sur les groupes H1 et H2 et les classes restantes. Les caractéristiques utilisées sont {EP2} et qmf328 ; $\dot{T}_M=0.9039$.

clc	111	H1	221	H2	411	511	512
111	98				2		
H1	1	93		2	1	2	1
221			100				
H2		17		76	3	4	
411		8		4	89		
511	2					95	3
512		3				3	94

Les indicateurs de bonne classification \dot{T}_M et κ pour le regroupement grossier (respectivement 0.90 et 0.69) sont meilleurs que lorsque les groupes sont plus dissociés (0.77 et 0.59), preuve de l’existence de similarités assez fortes entre les classes de chaque hyperclasse. Notons aussi que 17% des pixels de H2 sont classifiés dans H1. De même, dans le tableau A.3, tandis que les autres classes sont assez bien identifiées, le groupe {112 + 121 + 142} est le plus mal reconnu avec un taux de 53% seulement. Ceci s’explique par le fait que les classes de type *territoires artificialisés* qui sont plus du type “land use” que “land

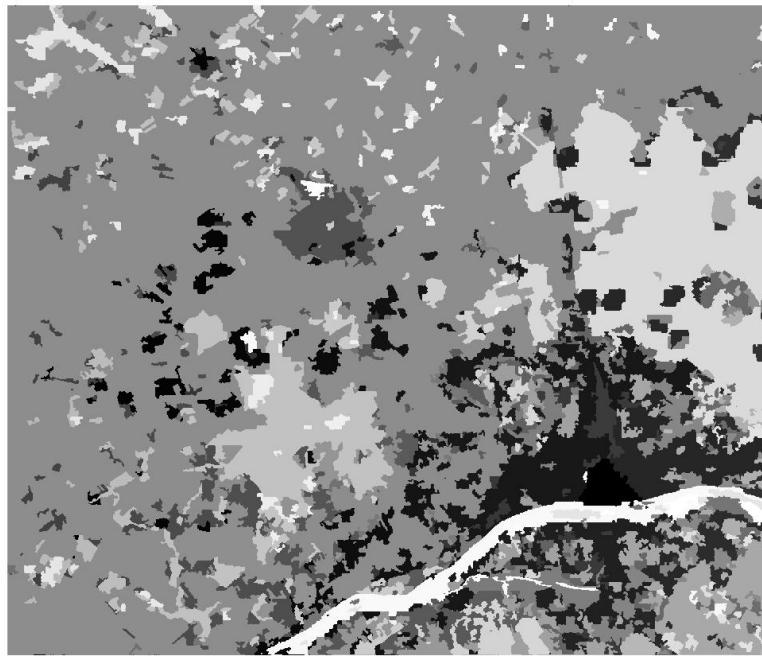


FIG. 4.11 – *Région 1* : résultats de la classification avec regroupement de classes. Certaines régions présentes dans la classification sont absentes dans la carte, illustrant ainsi une règle de généralisation de CLC.

cover”, sont très proches (on retrouve dans chacune d’elles, un peu d’espaces verts, des routes, du bâti) et se distinguent en général par le contexte. Ainsi, le groupe {112 + 121 + 142} est réparti dans 122 (*Réseaux routiers et ferroviaires*) à 8%, dans 124 (*Aéroports*) à 10%, dans 141 (*Espaces verts urbains*) à 8%, mais aussi dans le groupe des terres arables {211 + 231 + 242 + 243}.

Une autre cause de ce faible taux de classification est que le modèle considéré pour chaque groupe est celui de la classe qui a tendance à attirer les autres. Dans le cas de {112 + 121 + 142}, le modèle choisi est celui de la classe 121 (*Zones industrielles et commerciales*). Ceci veut dire que les pixels initialement bien classifiés dans 112 et 142 ont une grande probabilité d’être mal classifiés. Il serait alors intéressant de trouver un modèle qui décrirait assez bien les différents groupes, en considérant chaque hyperclasse comme un mélange de gaussiennes. L’algorithme EM (*Expectation Maximisation*) [Dempster et al., 1977] permet de mettre cette technique en œuvre.

Les figures 4.6 et 4.11 montrent respectivement la carte CLC et l’image classifiée obtenue après le regroupement. On note la présence de plusieurs régions qui ont disparu avec les règles de généralisation de CORINE Land Cover. Nous avons essayé de traiter ce problème avec les graphes d’adjacence.

4.5.3 Graphes d’adjacence

Pour la prise en compte des zones de mélange, nous avons procédé à une représentation de la zone par graphe d’adjacence [Tréneau & Colantoni, 2000], après une première

classification, puis à un traitement structurel du graphe en détectant les petites zones candidates à des regroupements, c'est-à-dire les régions ne respectant pas le critère de superficie minimale de CORINE Land Cover. Ces régions ont été associées à l'une de leurs voisines suivant un critère de distance minimale (distance euclidienne, distance de Mahalanobis). Il en est ressorti que la seule distance n'était pas suffisante pour retrouver les classes mélangées de CORINE Land Cover, et que le critère utilisé doit absolument tenir compte de la catégorie des classes voisines, et donc du contexte. Il existe des règles de priorité lorsque deux ou plusieurs grandes régions sont en conflit pour absorber une petite, cependant ces règles sont propres à chaque pays car établies selon la végétation et le type du paysage.

4.6 Applications sur un terrain inconnu

Nous testons l'apprentissage, non plus sur les données d'apprentissage, mais sur une autre région test, ceci pour mieux évaluer les capacités des méthodes que nous avons développées, à reproduire des cartes de type CORINE Land Cover sur des terrains inconnus.

Les données d'apprentissage et de test étant différentes, nous avons étudié deux cas : celui où les deux types de données appartiennent à la même scène, et celui où les deux types de données appartiennent à des scènes différentes. Seul le premier cas est présenté ici. Le deuxième est détaillé dans la section A.4 de l'annexe A.

Considérons la *région 1*. Ici, pour chaque classe, l'apprentissage s'est fait sur 80% des données choisies de manière aléatoire, et les tests, sur les 20% d'échantillons restants. La procédure suivie pour la classification est la même que précédemment.

Les tests ont été effectués en combinant les caractéristiques spectrales et texturelles et en tenant compte de l'information spatiale. Comme l'indique le tableau A.4 de l'annexe A présentant la matrice de confusion et la trace pondérée obtenues avec les primitives {EP2} et qmf328, les résultats sont assez similaires à ceux de la section précédente, où les tests étaient effectués sur les données d'apprentissage.

En effet, le tableau 4.15 récapitulant les traces pondérées obtenues dans les deux cas de figure pour chaque ensemble de caractéristiques, montre que les résultats des tests sur des données autres que celles d'apprentissage restent proches du cas où les deux types de données sont identiques. De plus, l'évolution des classes est la même, c'est-à-dire que les taux de reconnaissance de chaque classe diffèrent peu et restent du même ordre de grandeur. Par exemple, avec les caractéristiques {EP2}+ qmf328, la classe 111 (*Tissu urbain continu*) présente respectivement 97% et 96% de bonne classification.

TAB. 4.15 – Traces pondérées des matrices de confusion obtenues lors des tests sur les données d'apprentissage et sur des données différentes appartenant à la même scène.

	{EP1}+ qmf328	{EP2}+ qmf328	{EP3}+ qmf328
Données d'apprentissage ≡ données de tests	0.6257	0.6413	0.6007
Données d'apprentissage ≠ données de tests	0.6000	0.6378	0.5802

De même, le regroupement de régions tel que décrit dans la section 4.5.2, donne les résultats du tableau A.5 en annexe A, qui varient certes pour certaines classes, mais globalement restent proches des résultats du tableau A.3 lorsque les tests étaient effectués sur les données d'apprentissage. En effet, pour la même combinaison de regroupements, les indicateurs de bonne classification \bar{T}_M et κ pour le cas des données d'apprentissage et de tests différents (0.74 et 0.55 respectivement) ne sont pas très éloignés de ceux présentés par les tests sur les données d'apprentissage (0.77 et 0.59).

De manière individuelle, tandis que le taux de reconnaissance de certaines classes varie peu ou pas du tout (111, 122, 221, 331, 512, ...), d'autres classes sont mieux identifiées ; c'est le cas des *espaces verts urbains* (classe 141) qui a un taux de 86% (par rapport à 82% lorsque les données d'apprentissage et de test sont identiques), et le groupe {112, 121 et 142} dont le taux augmente de 4%. Par contre, le groupe {222, 311, 312, 313 et 324} n'est reconnu qu'à 71% (contre 82% lorsque les données d'apprentissage et de test sont identiques). Ces différences étaient prévisibles et sont en partie reliées à l'hypothèse de stationnarité des processus aléatoires (voir section 4.2.2).

L'application de l'apprentissage sur des données inconnues de la même scène donne des résultats proches et en accord avec ceux obtenus lors des tests sur les données d'apprentissage.

4.7 Conclusions

Cette étude que nous avons menée apporte des enseignements nombreux, enrichissant ainsi notre connaissance du domaine de la construction automatique de cartes thématiques d'occupation du sol.

Remarques générales Nous souhaitions élaborer une méthode dans laquelle quelques couples {images SPOT XS + cartes CORINE Land Cover} sont utilisés pour faire l'apprentissage des caractéristiques discriminantes des classes ; les connaissances ainsi acquises étant ensuite utilisées pour proposer des cartes Corine à partir des seules images SPOT XS. De cette étude, plusieurs points ont été mis en évidence.

Nous avons tout d'abord confirmé ce que l'évidence laissait prévoir : un petit nombre de classes ne seront pas aisément accessibles par des techniques de classification seules (même si elles utilisent des primitives avancées comme la texture où la morphologie des régions) car elles font référence à un usage des sols.

Par ailleurs, nous avons montré qu'il était possible, au sein d'une même image, de généraliser très efficacement une cartographie connue à partir de quelques zones d'apprentissage seulement. Pour cela, nous avons dégagé des primitives importantes radiométriques, et texturales et montré qu'elles étaient toutes les deux indispensables.

Mais surtout, nous avons montré comment il était souhaitable de combiner les classes et comment il fallait assembler les tests pour arriver à de tels résultats de façon hiérarchique en soulignant le rôle particulier joué par quelques classes très majoritaires qui absorbent les moins peuplées.

De plus, nous avons mis en évidence le fait que cette démarche était relativement robuste et pouvait se faire à partir d'un petit ensemble d'apprentissage à condition que la généralisation se fasse sur la même scène. En effet, même avec des images corrigées radiométriquement des dérives des capteurs, il n'est pas possible de transposer des tests

radiométriques multispectraux d'une image SPOT XS à une autre image, diachroniquement peu différente. Si l'on souhaite faire ce transfert, il nous semble indispensable de repasser à des images de réflectance des sols, corrigéant donc les variations de conditions atmosphériques.

Nous insistons sur le fait que les taux d'erreurs que nous mesurons ne sont pas significatifs. En effet, Nous mesurons le taux d'erreur en comparant nos classifications aux classes obtenues par CLC, mais les spécifications de CLC ne nous garantissent pas qu'une surface est pure à plus de 75 %. Rien ne nous permet ainsi de savoir quel est le taux exact de reconnaissance de nos algorithmes et nous ne pouvons conclure que sur la similarité de nos résultats avec ceux de CLC. Il se peut qu'ils soient nettement meilleurs, ... ou bien plus mauvais.

A propos des classes fonctionnelles Comme nous l'avons dit, les classes fonctionnelles telles que les *réseaux routiers et ferroviaires et espaces associés* ou les *aéroports*, posent des problèmes particuliers. Les critères radiométriques sont impropre à les mettre en évidence, ainsi que les critères texturaux. Il serait possible d'extraire des primitives structurelles (contours, régions, parallèles, etc.) qui seraient ajoutées aux primitives de radiométrie et de texture. Elles alourdiraient notablement le traitement et la classification, sans garantir des performances plus grandes car rappelons que les images SPOT utilisées ont une résolution de 20m. Les techniques plus classiques (par exemple utilisant des descripteurs locaux autour de points caractéristiques, comme les SIFT) semblent mal adaptées sur des images basses résolutions comme SPOT 2 XS. Elles sont pourtant capables de bonnes performances sur les éléments structurés d'un paysage, comme les pistes et taxi-ways d'un aéroport.

Le cas des classes de mélange Dans CLC, des règles strictes existent sur les tailles minimales des zones représentées. De plus des règles strictes existent sur la façon d'assembler des zones élémentaires ne vérifiant pas les règles ci-dessus, pour obtenir des zones conformes. Ceci peut nous amener à nous interroger sur l'intérêt d'une étude au niveau de la région, et non plus au niveau pixellaire comme nous l'avons fait jusqu'à présent. En effet, une classification basée sur le pixel peinera à retrouver ce type de classes si elle n'est pas suivie d'un traitement visant à regrouper, probablement en utilisant des critères en plus (géométriques par exemple), de petites zones connexes candidates pour une classe de mélange. Cependant, s'il semble probable que nous trouverons des mélanges dont les proportions répondent bien à celles des exemples appris dans CLC, il est peu probable que nous aurons appris les véritables critères qui président au choix des frontières des classes de mélange. Il est en effet probable que les critères adoptés par le cartographe relèvent de décisions géomorphologiques (appartenance à un même bassin versant par exemple) ou à des critères administratifs (limites de départements) qui ne seront pas reconnus par l'apprentissage, d'où l'importance de l'information exogène dans CORINE Land Cover.

Dans la suite de ce document, nous nous focalisons sur ce problème tangible, et proposons de nous appuyer sur une approche par régions pour aborder le problème de l'identification des classes de mélange à partir d'une image satellitaire.

Chapitre 5

Représentation des images basée sur une approche par régions

Les études menées dans le chapitre 4 mettent en évidence les limitations des approches classiques pour l'identification de classes dites de mélange. La difficulté d'apprendre les classes de mélange, telles que les *forêts mélangées* dans la classification CORINE Land Cover (CLC), et de les reconnaître dans de nouvelles images est un obstacle majeur pour la production automatique des cartes d'occupation du sol. En effet, ces classes sémantiques qui sont des mélanges ou regroupements d'autres types d'occupation des terres, sont très présentes dans les images satellitaires, qu'elles soient à haute ou à basse résolution. Par exemple, nous avons identifié dans des images Quickbird de Las Vegas à 61 cm de résolution, la classe des *banlieues résidentielles*, qui est principalement composée de maisons, d'espaces verts, de terrains nus et de piscines (voir figure 5.1). Ainsi, comme pour les *forêts mélangées* ou les *systèmes culturaux et parcellaires complexes* dans CLC, une classification basée sur une approche classique ne sera pas suffisante pour identifier correctement les *banlieues résidentielles*.

Forts des conclusions tirées des expérimentations effectuées dans le chapitre 4, nous remettons en cause l'étude basée sur une approche au niveau du pixel et proposons donc dans ce chapitre, de nous appuyer sur une approche par régions pour déterminer les classes de mélange dans les images satellitaires. Contrairement aux méthodes de classification basées sur une approche par régions présentées dans l'état de l'art, nous ne souhaitons pas classifier chaque région. L'idée ici, est de représenter les images sous une forme particulière basée sur les régions, de manière à profiter des propriétés de chacune d'elles, mais aussi des relations existant entre elles.

5.1 Potentiel de l'approche par régions

L'intérêt d'une telle approche est multiple. Outre le fait que le partitionnement de l'image en régions tient déjà compte de l'information spatiale, et permet de manipuler des régions homogènes, elle réduit aussi de manière considérable la taille des données. En outre, suivant la qualité du partitionnement, elle peut faire émerger des régions saillantes ou des objets de l'image.

Mais surtout, l'intérêt de cette approche, en ce qui nous concerne, est que, présentée sous une certaine forme, elle permet l'utilisation des méthodes de traitement des données

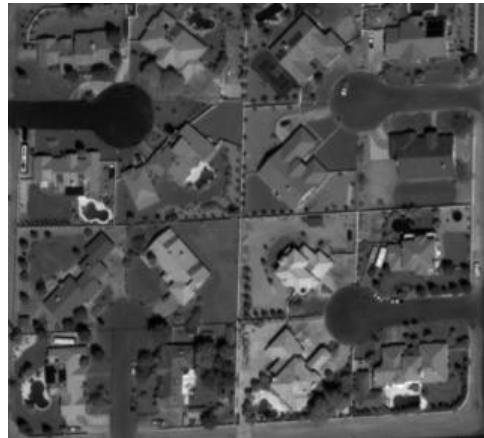


FIG. 5.1 – Image Quickbird de Las Vegas à 61 cm de résolution, appartenant à la classe des *banlieues résidentielles*. C'est une classe de mélange, constituée principalement de maisons, d'espaces verts, de terrains nus et de piscines.

textuelles exprimées dans une langue. Nous nous intéressons ici aux notions de “compositionnalité sémantique” et de “représentation en sacs-de-mots” du langage textuel.

5.1.1 Principe de compositionnalité sémantique

Nous souhaitons exploiter le “principe de compositionnalité sémantique” [Pelletier, 1994], et l’appliquer aux images satellites. Très utilisé en traitement automatique des langues (TAL), le “principe de compositionnalité sémantique”, encore appelé “principe de Frege”, spécifie que la signification d’une expression complexe est fonction de la signification de ses parties et de leurs relations syntaxiques. En effet, le principe de compositionnalité présente une démarche très utilisée pour le traitement automatique de la sémantique. Il permet de formuler un ensemble de règles de combinaison pour dériver le sens d’un énoncé à partir de celui de ses composantes. Prenons l’exemple de la phrase suivante, présentée dans [Bouillon & Vandooren, 1998] : “Le petit chat que Marie a acheté s’est endormi en ronronnant”. Le sens de cette phrase contient bien, entre autres, les sens de *petit*, *chat*, *acheté*, *endormi*, *ronronnant*, et les liens entre eux.

Ainsi, de même que la compréhension d’une phrase nécessite de comprendre les mots et la syntaxe entre ces mots, nous nous proposons d’exploiter l’information spatiale existant entre les régions d’une image satellite, afin d’analyser et décrire son contenu sémantique.

Cependant, l’application de ce principe n’est pas toujours simple, car il ne se vérifie pas pour toutes les expressions de la langue naturelle (expressions idiomatiques par exemple). De plus, le principe de compositionnalité est compliqué par le fait que les relations entre les structures syntaxique et sémantique d’un énoncé ne sont pas toujours biunivoques : à une seule structure syntaxique, peut correspondre plus d’une représentation sémantique et vice versa. En ce qui concerne les images, et en particulier les images satellites, il nous paraît plus naturel de supposer qu’à chaque unité syntaxique, correspond une règle d’interprétation sémantique unique, par exemple, une *maison* et une *piscine* correspondent à une *villa* : c’est l’hypothèse forte du principe de compositionnalité (*rule-to-rule hypothesis*). Par conséquent, nous ne tiendrons compte que du sens propre

des images.

5.1.2 Approche par sacs-de-mots

Nous proposons par ailleurs, d'utiliser les techniques basées sur une représentation dite en "sacs-de-mots" (*bags-of-words*) pour classifier les images satellitaires, et en particulier identifier les classes complexes telles que celles de mélange. Le modèle par sac-de-mots est devenu particulièrement populaire durant ces dernières années, en raison de la qualité des résultats qu'il permet d'obtenir dans le domaine de la fouille de données dans les textes. Ce type de représentation, très utilisée en analyse de texte, suppose que l'ordre des mots dans un document peut être négligé. Cela consiste à décrire chaque document au moyen d'un histogramme des occurrences de chaque mot du vocabulaire. L'histogramme, plus ou moins pondéré par la fréquence d'apparition des mots dans toute la langue, est ensuite utilisé comme vecteur de forme par un algorithme de classification, utilisant des modèles génératifs ou la classification par recherche de fonctions discriminantes.

Cette approche a été adaptée pour l'annotation automatique et la recherche d'images multimédia par le contenu sémantique [Monay & Gatica-Perez, 2003; Fei-Fei & Perona, 2005; Lazebnik et al., 2006; Larlus & Jurie, 2008].

Nous nous intéressons ici, à l'exploitation d'une telle approche pour l'annotation de grandes images satellitaires.

L'utilisation des techniques textuelles pour les images, que ce soit pour l'exploitation du principe de compositionnalité sémantique ou de l'approche par sacs-de-mots, nécessite que l'image soit décrite sous forme de "mots visuels", par analogie au texte. Cependant, les images n'ont pas de vocabulaire a priori, le "vocabulaire visuel" doit donc être construit pour répondre à des attentes particulières. La section suivante est consacrée à la définition des mots visuels, et à la construction du vocabulaire.

5.2 Codage de l'image : les mots visuels

5.2.1 Le vocabulaire visuel

Le vocabulaire visuel est défini par analogie avec le vocabulaire de mots textuels, dans le but de pouvoir appliquer les techniques statistiques de textes aux images. Aucun vocabulaire visuel n'existant de manière explicite, il faut donc le construire de manière à ce qu'il représente au mieux les données. Les mots visuels sont obtenus par quantification vectorielle de descripteurs locaux extraits des images : il s'agit d'une transformation de l'espace de description vers un espace discret d'étiquettes. La représentation de l'image par des mots visuels peut donc être vue comme une sorte de codage de l'image. De manière générale, un codage permet de passer d'une représentation des données à une autre. En compression des données (théorie de l'information), l'information à compresser est vue comme la sortie d'une source de symboles qui produit des textes finis selon certaines règles et le but est de minimiser la taille moyenne des textes obtenus en réduisant l'information répétitive, c'est-à-dire la redondance. En effet, la compression des données permet de réduire l'espace nécessaire à la représentation d'une certaine quantité d'information. Par analogie, le codage de l'image consiste à transformer l'image en une représentation plus concise et facilement exploitable par l'utilisateur, tout en préservant son contenu informationnel.

La technique la plus utilisée pour construire un vocabulaire visuel consiste à répartir dans des groupes (*clusters*), des descripteurs locaux extraits des images d'apprentissage, le nombre de clusters représentant la taille du vocabulaire. La construction d'un vocabulaire visuel dépend donc fortement du type de primitives extraites localement des images, et de l'algorithme utilisé pour la quantification.

5.2.2 Etat de l'art

Etant donné un ensemble d'images, la méthode la plus courante de codage est celle utilisée dans [Barnard et al., 2003; Duygulu et al., 2002; Jeon et al., 2003]. Elle consiste à segmenter les images en régions, et à extraire pour chaque région, un vecteur de caractéristiques (spectrales, texturelles, formes, etc.) qui la représente. Ces représentations des régions sont ensuite quantifiées par un algorithme de clustering (en général les *k-means*). Les groupes de régions (*clusters*) obtenus, que nous appellerons *blobs* comme dans [Jeon et al., 2003; Barnard et al., 2003; Duygulu et al., 2002], sont représentés chacun par un "mot visuel", qui est le centroïde du cluster. Ainsi, chaque blob a une étiquette qui permet de l'identifier. Les étiquettes sont assimilées à un dictionnaire visuel pouvant être utilisé pour décrire le contenu visuel de l'image. Une variante de cette méthode de codage est de découper l'image en imagettes en utilisant une grille régulière, plutôt que de la segmenter. Cette dernière technique est utilisée dans les travaux de Mori et ses collègues [Mori et al., 1999] et Li et Bretschneider [Li & Bretschneider, 2006].

Par ailleurs, au lieu de traiter l'image de manière dense, il est possible de sélectionner un sous ensemble de points qui sont plus informatifs que les autres, à l'aide d'un critère de saillance, réduisant ainsi la quantité d'information à traiter. Par exemple, Fei-Fei et Perona [Fei-Fei & Perona, 2005], dans leur approche d'identification de la sémantique des scènes naturelles, construisent un vocabulaire visuel à partir d'une base d'apprentissage sur lesquelles sont extraites des régions d'intérêt, obtenues de 4 manières différentes (entre autres le détecteur de Kadir et Brady). Les descripteurs locaux SIFT [Lowe, 2004] sont calculés sur ces régions, puis regroupés en clusters avec les *k-means* pour obtenir les mots du vocabulaire visuel, dénués de sémantique, mais constituant une description pertinente des images (voir figure 5.2). De même, dans [Marszaek & Schmid, 2006], les auteurs construisent un vocabulaire visuel en extrayant des descripteurs SIFT de points d'intérêt repérés avec les détecteurs de Harris-Laplace et Laplacien, puis en les quantifiant à l'aide des *k-means*. Leibe et ses collègues [Leibe et al., 2006] quant à eux, utilisent plutôt un algorithme de quantification hiérarchique pour regrouper les descripteurs.

D'autre part, une autre méthode, moins classique, de codage d'images couleur est détaillée dans [Watanabe et al., 2002] et propose un schéma de représentation de l'image basé sur la compression de données. L'image est segmentée puis chaque région est encodée en un texte : chaque région est tout d'abord représentée par un graphe pondéré dans lequel les noeuds sont les pixels de la région et le poids d'un arc est donné par la différence spectrale entre les deux noeuds reliés ; puis, l'arbre recouvrant de poids minimal de ce graphe est extrait et parcouru selon une certaine direction. Une quantification vectorielle appliquée sur les couleurs des noeuds et les directions de parcours permet d'obtenir le texte code. Ensuite, un ensemble de compresseurs de texte est utilisé pour générer un vecteur de taux de compression qui servira de caractéristique pour la région. Il est aussi possible de travailler directement sur toute l'image pour obtenir un vecteur de taux de compression caractérisant l'image.

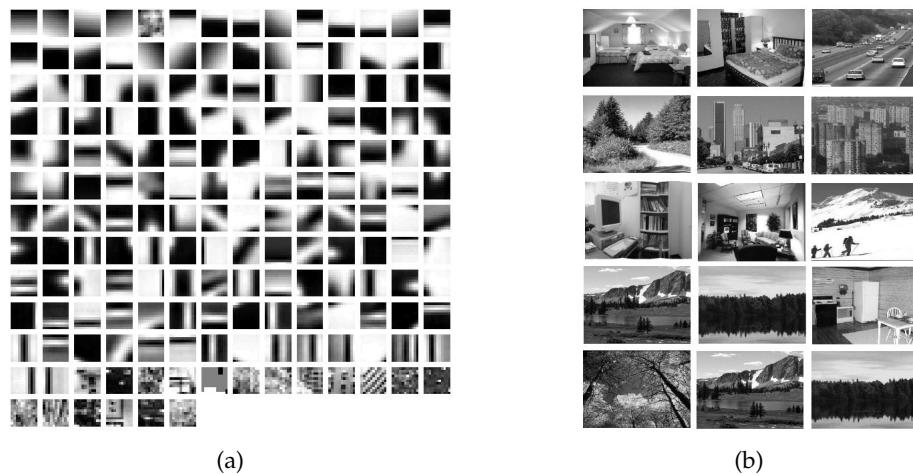


FIG. 5.2 – (a) Exemple de mots du vocabulaire visuel utilisé par Fei-Fei et Perona. Ces mots visuels, qui sont des fenêtres de pixels, sont issus d’images naturelles (b) appartenant à 13 concepts sémantiques, tels que chambres, salons, bureau, paysages, forêts ou encore autoroutes.

Une fois l’image caractérisée par l’une des méthodes que nous venons de décrire, en fonction de l’application, différentes techniques (généralement d’apprentissage) sont utilisées pour introduire une sémantique. En effet, dans la littérature, outre les tâches de compression pour la transmission, le codage d’une image a par ailleurs été utilisé pour interpréter sémantiquement des scènes ou rechercher des images par mots-clés : on parle alors de “codage sémantique”. C’est une étape nécessaire dans le processus d’analyse de l’image pour des applications telles que l’annotation, la correspondance ou la reconnaissance d’objets dans les images naturelles. Notons que l’opération d’annotation est une forme de codage sémantique pour lequel les mots sont directement les symboles utilisés pour le codage.

5.2.3 Approche choisie

La question de la construction d’un vocabulaire visuel pertinent est donc essentielle, car de l’aptitude des mots visuels à bien représenter les images de l’ensemble d’apprentissage, dépendent les performances des traitements ultérieurs (catégorisation, annotation, etc.).

Nos travaux concernent les images satellitaires, qui en général sont riches en information. Afin de supprimer tout risque de manquer des informations utiles à la classification, certains travaux considèrent tous les pixels de l’image sans aucune sélection [Winn et al., 2005], ou alors échantillonnent les images aussi densément que possible. Cependant ces méthodes traitent chaque pixel séparément, ce qui n’est pas notre objectif. Nous avons donc considéré deux types d’approches :

- Les images sont découpées suivant une grille régulière dont la taille dépend de la résolution de l’image et du type de classes sémantiques recherchées. Les mots visuels seront donc calculés à partir de fenêtres de pixels.
- Les images sont partitionnées en régions à l’aide d’un algorithme de segmentation. L’étape de segmentation est essentielle, car elle est irréversible : une mauvaise seg-

mentation se répercute en cascade sur la qualité des traitements ultérieurs. Ici, nous avons utilisé l'algorithme Mean shift présenté dans la section 3.1, qui en permettant de fixer la taille de la région minimale, permet de contrôler le nombre de régions de l'image segmentée. Nous évitons ainsi le nombre important de petites régions non significatives qu'on obtient avec l'algorithme watershed par exemple. L'évaluation de la qualité de la segmentation est visuelle. Il s'agit de voir si les régions de l'image segmentée correspondent plus ou moins à certains objets de l'image ou à des parties évidentes d'un objet. Une image légèrement sursegmentée ne serait donc pas problématique. Proposer une évaluation objective de la qualité des segmentations produites n'est pas simple, nous pourrions les comparer par exemple à l'aide de critères entropiques [Meila, 2002].

Les primitives extraites des images dépendent essentiellement du type et de la résolution des images. L'étude effectuée dans le chapitre 4 sur CORINE Land Cover, à l'aide des images SPOT2 XS a montré que pour ces images, la combinaison des caractéristiques spectrales, y compris les néocanaux, et texturelles donne en général de meilleurs résultats. Dans le chapitre 7, lors de l'annotation des images Quickbird panchromatiques de Las Vegas à 60 cm de résolution, nous avons utilisé la moyenne et la variance comme descripteurs des fenêtres de pixels issues du découpage suivant une grille régulière, cependant les coefficients de l'analyse en composantes principales (ACP) sont une alternative. Lorsque les images seront segmentées au préalable, nous extrairons en plus pour chaque région, des caractéristiques de forme que nous définirons lors de leur utilisation.

L'étape de quantification vectorielle est basée sur l'algorithme de clustering k -means. Cependant, un paramètre important à déterminer est le nombre de clusters, qui représente la taille du vocabulaire. Afin de déterminer le nombre optimal de clusters, nous utilisons le critère KMDL (*kernel MDL*) décrit dans [Kyrgyzov et al., 2007], qui modélise les descripteurs par un mélange de gaussiennes, et utilise le critère de la longueur de description minimale (*Minimal Description Length* ou MDL) pour accéder à la complexité optimale du modèle. De cette manière, les centroïdes des clusters produits en utilisant les k -means et le nombre optimal de clusters obtenu, constituent les mots visuels du vocabulaire.

Chaque région ou fenêtre de pixels de l'image est ainsi assimilée au mot visuel qui lui est associé, et a donc une étiquette.

5.3 Relations spatiales et représentation

Une fois l'image décomposée en mots visuels, la détermination de la sémantique de l'image peut être facilitée par une étape de prise en compte de relations syntaxiques, suivant le principe de compositionnalité sémantique. En effet, les régions élémentaires peuvent être regroupées de manière à prendre en compte les relations spatiales et contextuelles, pour former des structures sémantiques.

5.3.1 Analyse syntaxique

Il est possible d'exploiter les relations syntaxiques entre les mots visuels pour déterminer la sémantique de l'image. Par exemple, des méthodes faisant usage de réseaux bayésiens exploitent les relations entre les régions de l'image issues d'une segmentation. Dans

leur méthode contextuelle par boucle de pertinence [Li & Bretschneider, 2006], Li et Bretschneider modélisent les liens entre les codes des paires de régions adjacentes (mots visuels) de l'image et les concepts sémantiques, pour calculer des fonctions de score sémantique. Ces dernières mesurent l'importance du lien entre une image ou une région, et un concept sémantique en se basant sur les matrices de cooccurrence des codes. Aksoy et ses collègues [Aksoy et al., 2005] utilisent une modélisation floue des relations spatiales entre des paires de régions pour décrire des concepts de haut niveau (disjoints, proche, droite, etc). Dans le domaine du texte, les relations spatiales sont introduites par exemple en comptant les occurrences des paires de mots dans un texte.

Dans cet ordre d'idées, nous nous intéressons donc à l'analyse des relations spatiales entre les mots du vocabulaire visuel pour mettre en évidence, les structures d'intérêt de l'image. Les relations spatiales constituent un élément essentiel des descriptions d'agencement spatial entre les régions d'une scène et sont donc utiles à un grand nombre de tâches liées à la reconnaissance des formes, la vision par ordinateur, les systèmes d'information géographique (SIG), et plus particulièrement l'interprétation des scènes. En effet, nous utilisons considérablement les relations spatiales pour décrire, détecter et reconnaître les objets d'une scène. Par exemple, dans leur architecture multi-spécialistes pour l'interprétation d'images MESSIE [Giraudon et al., 1992], Giraudon, Garnesson et Montésinos proposent de modéliser les objets physiques d'une scène, en se basant sur une organisation selon 4 points de vue : la forme géométrique de l'objet, sa radiométrie, ses relations spatiales avec d'autres objets et sa fonctionnalité. Ils traduisent les relations spatiales par une liste "d'heuristiques de localisation" qui expriment des relations topologiques entre différents objets ou parties d'objets, par exemple, une ombre est à côté d'un bâtiment, un pont est sur une rivière, etc. Cette prise en compte du contexte permet, soit de valider un objet par l'existence d'autres objets déjà détectés dans son contexte réel, soit d'inférer une nouvelle hypothèse d'objet.

La description de l'agencement spatial peut varier très significativement d'une langue à l'autre [Landau & Jackendoff, 1996]. Freeman [Freeman, 1975] a étudié l'ensemble de ces relations pour la langue anglaise et a proposé une liste de treize relations spatiales de base pour décrire la position relative de deux objets dans un espace bidimensionnel. Dans la littérature, ces relations sont généralement classées en trois familles : relations topologiques, relations métriques (ou distances) et directions. On rajoute souvent à ces trois familles les relations morphologiques ou encore les relations de symétrie.

Les relations topologiques sont les relations de contact et de connexité entre deux régions. Elles font appel à des notions de la théorie des ensembles (à l'intérieur de, à l'extérieur de,) ou des notions de voisinage (adjacence). Les relations métriques sont celles qui ont un lien direct avec la distance qui sépare les deux objets concernés (proche de, loin de). Les relations directionnelles, quant à elles, regroupent l'ensemble des relations faisant appel à une direction de l'espace ou du plan (à droite de, à gauche de, au-dessus de, en-dessous de, devant, derrière).

5.3.2 Représentation des relations spatiales

Pour représenter ces relations spatiales, deux types de modèles sont utilisés dans la littérature : les modèles de représentation qualitative souvent liés à la logique formelle ou à la classification des intervalles temporels [Zahzah et al., 2003], et les modèles de représentation quantitative plutôt rattachés à la logique floue ou aux probabilités [Bloch, 2005]. La première catégorie de modèles a été développée pour représenter les rela-

tions spatiales topologiques [Mark & Egenhofer, 1994; Clementini & Felice, 1997]. On y retrouve aussi les modèles reposant sur la théorie de la méréotopologie. Par ailleurs, il existe quelques approches qualitatives des relations directionnelles [Freksa & Zimmerman, 1992]. Cependant, certains résultats obtenus via cette approche sont imprécis, plusieurs configurations étant possibles.

L'ambiguïté liée à la définition des relations spatiales a encouragé plusieurs chercheurs à adopter les relations floues introduites par Zadeh [Zadeh, 1965]. Le formalisme flou permet de prendre en compte l'imprécision et l'incertitude attachées d'une part aux caractéristiques des relations elles-mêmes (certaines relations comme les directions sont intrinsèquement imprécises), et d'autre part, aux imperfections des données issues des images. De plus, cette démarche fournit une évaluation numérique de la satisfaction des relations dans une image donnée (degré de satisfaction et non plus relation binaire). On peut distinguer deux types d'approches floues pour la représentation des relations spatiales : soit on définit une valeur représentant le degré de satisfaction des relations entre deux objets donnés, soit on définit un ensemble flou représentant en tous points de l'espace, le degré de satisfaction de la relation par rapport à un objet de référence. Aksoy et ses collègues [Aksoy et al., 2005] se basent sur la première approche et utilisent une modélisation floue des relations spatiales entre des paires de régions pour décrire des concepts de haut niveau (disjoints, proche, droite, etc).

Contrairement aux relations topologiques, les relations de distance ont essentiellement été traitées par des approches quantitatives. Il est cependant possible de définir une relation floue à partir d'une relation binaire, ce qui donne naissance par exemple à l'adjacence floue [Bloch et al., 1997].

Dans notre étude, nous nous intéressons aux relations topologiques et plus particulièrement, à l'adjacence. En effet, les relations spatiales sont étudiées entre deux régions voisines, c'est-à-dire que deux régions élémentaires ne peuvent être regroupées que si elles sont adjacentes. Nous exploitons la notion d'adjacence floue dans le sens où un degré d'adjacence est défini entre deux régions voisines de l'image. Ce degré d'adjacence est en général une fonction décroissante de la distance entre les deux objets considérés. Dans notre cas, elle dépendra de paramètres géométriques tels que la forme ou la compacité des régions ainsi que de la taille relative de la frontière commune entre les deux régions. Ces critères ont leur influence dans la prise de décision quant au regroupement d'une région élémentaire avec une autre.

5.4 Apprentissage

Les mots visuels sont largement utilisés dans les méthodes d'apprentissage appliquées aux images, basés sur une représentation en sacs-de-mots. Cette représentation d'images suppose la non-pertinence de l'ordre d'apparition des mots dans un document : les images sont donc vues comme des collections de mots visuels au même titre que les textes sont vus comme des collections de mots. Les méthodes de classification de texte basées sur les mots deviennent donc applicables. En effet, la représentation en sacs-de-mots s'est avérée très efficace pour la classification, mais aussi l'annotation d'images. Sa principale force réside dans la représentation compacte associée à chaque image : un histogramme de fréquence des mots visuels dont la dimension est égale à la taille du vocabulaire visuel. Cette représentation est souple, et assez robuste à certaines variations des objets

de l'image. L'histogramme (qui peut être normalisé pour s'affranchir de la taille du document et pour donner plus d'importance aux mots discriminants et/ou fréquents) est ensuite utilisé comme vecteur de forme par un algorithme de classification.

Dans la littérature, les travaux de classification, d'annotation ou de recherche d'images multimédia par le contenu, utilisant une approche par sacs-de-mots sont nombreuses. Nous nous intéressons plus particulièrement à l'annotation automatique d'images, qui est un domaine du traitement d'images, permettant d'associer automatiquement des mots-clés ou du texte à des images, afin de pouvoir ensuite rechercher les images par requête textuelle.

5.4.1 Annotation automatique d'images

Parmi les méthodes d'annotation utilisant une représentation de l'image en sacs-de-mots, on distingue les annotations directes qui font directement le lien entre les caractéristiques bas-niveau et les annotations sémantiques, et les approches faisant usage d'une ou de plusieurs couches intermédiaires.

5.4.1.1 Annotation directe

Au lieu de chercher à attribuer un ou plusieurs mots-clés à l'image en la considérant comme un tout, la plupart de ces méthodes indiquent en général les correspondances entre les mots-clés et les parties de l'image, ce qui peut s'avérer plus intéressant, car plus précis. En effet, s'il est vrai que certains concepts peuvent concerner l'ensemble de l'image (intérieur, extérieur, etc), d'autres par contre sont plus spécifiques à une partie de l'image (fleur, tigre, ciel, etc). Dans ces approches, l'on s'intéresse à la probabilité d'associer les mots-clés aux régions de l'image appartenant aux différents blobs. Chaque région étant traitée indépendamment des autres, l'image est donc représentée comme un sac de mots.

Mori, Takahashi et Oka [Mori et al., 1999] proposent une méthode pour associer des mots aux régions de l'image, à partir d'images annotées globalement, ce qui permet de s'affranchir de l'étape laborieuse de constitution manuelle d'une base de régions. A l'aide d'une grille rectangulaire de taille évolutive, chaque image d'apprentissage annotée est divisée en blocs, qui héritent chacun des annotations de l'image globale. Une quantification vectorielle est ensuite appliquée sur les caractéristiques extraites de chaque bloc (histogramme RVB cubique ainsi que des histogrammes de direction des contours de Sobel), de manière à former des clusters ou *blobs*. C'est l'étape de codage de l'image. Un modèle de cooccurrence des mots m_i et des clusters c_j obtenus est alors appris en estimant la probabilité conditionnelle d'un mot sachant un cluster $P(m_i | c_j)$.

De même, les travaux menés par Duygulu [Duygulu et al., 2002] utilisent un "modèle de traduction" pour associer un mot à une région individuelle de l'image, à partir d'annotations attachées à l'image entière. Ce modèle se comporte comme un "lexique" qui, sachant les mots dans une langue, les prédirait dans une autre langue : la reconnaissance d'objets est assimilée à un système de traduction automatique, qui transformerait un vocabulaire de blobs en un vocabulaire de mots. Cette correspondance entre les régions de l'image et les mots peut être apprise grâce à une méthode basée sur EM (*Expectation Maximisation*). En effet, un ensemble d'apprentissage est utilisé pour construire une table (pouvant être apprise de manière itérative) représentant les probabilités conditionnelles d'un mot sachant une région. Etant donnée une nouvelle image dont les blobs ont été

identifiés, le mot ayant la plus forte probabilité sachant le blob est attribué aux régions contenues dans ce blob. Cette méthode a été approfondie et comparée à d'autres modèles dans [Barnard et al., 2003].

En outre, dans [Jeon et al., 2003], les images sont aussi décrites en utilisant un vocabulaire de blobs, puis un modèle de pertinence cross-media (*cross-media relevance model*) est utilisé pour estimer la probabilité conjointe de la présence d'un mot et d'une région dans une image donnée. Ce modèle a été adapté dans [Lavrenko et al., 2003] pour utiliser une fonction de densité de probabilité continue (CRM ou *continuous relevant model*), afin de décrire le processus de génération des caractéristiques à partir des régions, espérant éviter la perte d'information causée par la quantification.

Notons justement que les performances d'annotation des méthodes que nous venons de décrire dépendent énormément de la qualité du clustering effectué au niveau des régions. En effet, un nombre de clusters trop grand ou trop petit peut être pénalisant pour l'annotation. Pour remédier à ce problème, une approche proposée dans [Yang et al., 2005] consiste à utiliser le *Multiple Instance Learning*.

5.4.1.2 Annotation utilisant un niveau intermédiaire

Le fossé sémantique existant entre les caractéristiques bas-niveau et la sémantique de l'image a récemment poussé certains travaux de recherche vers l'utilisation d'une représentation sémantique intermédiaire pour réduire ce fossé, en utilisant la connaissance que nous avons des images. Cette approche, connue sous le nom de "modélisation sémantique" (*semantic modelling*), propose d'utiliser cette information supplémentaire pour améliorer les performances des classifieurs entraînés uniquement à partir de caractéristiques de bas-niveau. Pour la reconnaissance des scènes, l'étape intermédiaire peut être la reconnaissance des objets de l'image par analyse de régions, l'extraction de descripteurs locaux autour de points d'intérêt afin de faire la classification (identification des mots visuels, puis apprentissage), ou encore l'utilisation des propriétés sémantiques de l'image.

Dans le système KIM (*Knowledge-driven Information Mining*) [Datcu et al., 2003], un modèle hiérarchique à 5 niveaux est utilisé pour l'interprétation sémantique du contenu de l'image tel que défini par l'utilisateur, à partir des données D correspondant au niveau 0 (figure 5.3). Les niveaux 1 et 2 correspondent à l'extraction des caractéristiques et des métacaractéristiques. Des modèles stochastiques M sont utilisés pour capturer les structures spatiales, spectrales et géométriques de l'image. Ces modèles, paramétriques, sont décrits en termes de formalisme bayésien par $p(D | \theta, M)$, qui correspond à la probabilité d'une réalisation de D pour une valeur particulière du vecteur de paramètres θ . Le but est d'inférer le vecteur de caractéristiques θ , sachant les données D , à l'aide de différents estimateurs tels que le Maximum a Posteriori (MAP). A partir de cette modélisation bayésienne, les métacaractéristiques du niveau 2 sont générées. Ensuite, à partir des éléments obtenus aux niveaux 1 et 2, un vocabulaire de classes signal w_i est obtenu pour chaque modèle M , via une classification non supervisée. Notons que cette étape de caractérisation de l'image (niveaux 1 à 3) est complètement non supervisée, et donc objective. Alors, l'utilisateur entre en action et définit les concepts sémantiques qui l'intéressent. Ces concepts sémantiques L_ν sont liés à des combinaisons du vocabulaire de classes signal en utilisant des réseaux bayésiens : $p(w_i | L_\nu)$. Ces liens sont mis à jour grâce au retour de l'utilisateur pendant le processus d'apprentissage. De ce qui précède, la relation entre les données D et les classes sémantiques est obtenue par :

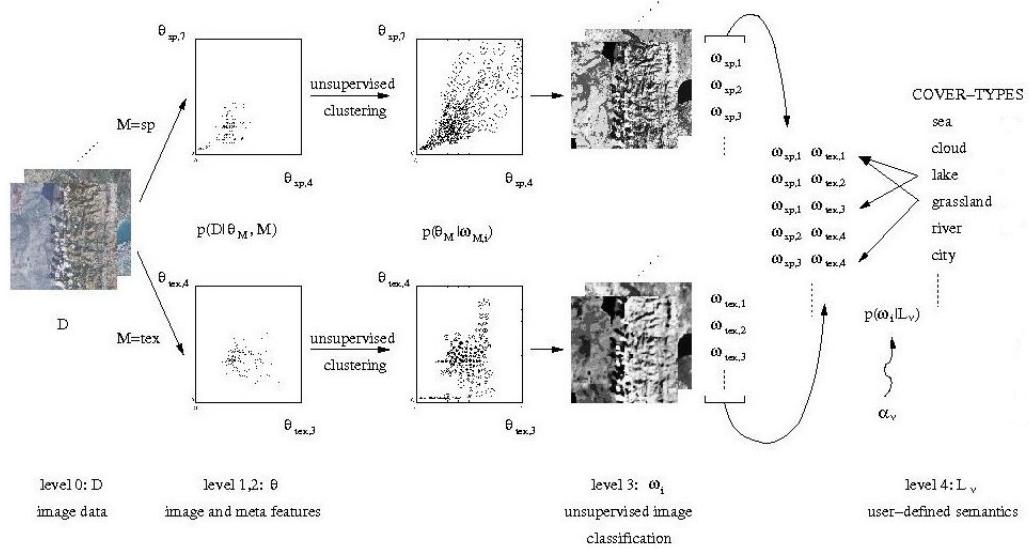


FIG. 5.3 – Description hiérarchique du contenu de l'image et de la sémantique de l'utilisateur utilisée dans le système KIM.

$$p(L_\nu | D) = \sum_i p(L_\nu | w_i) p(w_i | D). \quad (5.1)$$

Cette méthode est basée sur un apprentissage interactif (*Interactive learning*). En effet, quelques systèmes ont intégré des interfaces où les utilisateurs peuvent interagir avec le système, afin d'améliorer les résultats de l'annotation. Le système peut ainsi apprendre les intentions de l'utilisateur et rendre des résultats pouvant le satisfaire. Cet enrichissement de la requête se traduit par la boucle de pertinence (*Relevance Feedback*). Au début, les poids des descripteurs pour chaque image de la base sont fixes et objectifs, puisque calculés de façon indépendante, tandis que les requêtes de l'utilisateur sont subjectives par nature. L'objectif est de faire des interactions entre le système et l'utilisateur afin de faire refléter cette subjectivité dans le poids des descripteurs.

Une autre catégorie de ces approches est celle basée sur les modèles génératifs probabilistes à variable latente tels que la pLSA (*probabilistic Latent Semantic analysis*), le LDA (*Latent Dirichlet Allocation*) et leurs dérivés. Ces modèles fournissent une approche probabiliste de la modélisation de documents textuels comme un mélange d'aspects ou "topics" intermédiaires, émergeant du document de manière non supervisée. En effet, les modèles pLSA et LDA sont hiérarchiques et modélisent chaque mot d'un document comme étant émis par une variable latente, qui a une distribution dépendant de chaque document, et qui représente les topics présents dans celui-ci. Dans les images, les poids des topics latents dans le mélange indique la proportion de chaque objet ou structure dans l'image. Nous décrivons ces deux modèles plus en détail et discutons de leur utilisation pour la modélisation des images dans la section 5.4.2.

Dans cet ordre d'idées, Fei-Fei et Perona [Fei-Fei & Perona, 2005] proposent une approche d'identification de la sémantique des scènes naturelles, en se basant sur le modèle LDA. Les auteurs construisent un vocabulaire visuel, à partir d'une base d'apprentissage

sur lesquelles sont extraites des régions d'intérêt. Les descripteurs locaux SIFT [Lowe, 2004] sont calculés sur ces régions, puis regroupés en clusters pour obtenir les mots du vocabulaire visuel, dénués de sémantique, mais constituant une description pertinente des images. La description en sacs-de-mots de l'image est utilisée pour l'apprentissage de 13 catégories sémantiques : chambre, salon, autoroutes, grands bâtiments, banlieues, côte, cuisine, bureau, centre-villes, rues, forêts, montagnes et paysages. Les mots sont d'abord classés localement en différents topics intermédiaires, représentant des concepts sémantiques locaux, puis en catégories sémantiques.

Dans [Monay & Gatica-Perez, 2003], les auteurs utilisent le modèle pLSA pour annoter des images naturelles entières. Dans leur approche, les images sont tout d'abord segmentées en 3 régions fixées : le centre, la partie haute et la partie basse. Les histogrammes de couleur extraits de chaque région sont quantifiés, et rassemblés dans un sac-de-mots. Pour les images d'apprentissage qui sont annotées, les occurrences des termes de l'annotation sont rajoutées dans le sac de mots. Cette représentation de l'image est utilisée avec le modèle pLSA pour l'annotation des images, ainsi qu'avec sa version non probabiliste LSA (*Latent Semantic Analysis*) [Deerwester et al., 1990], basée sur une décomposition en valeurs singulières (SVD). Les auteurs montrent que les performances sont comparables, malgré le fait que modèle pLSA soit plus complexe que le modèle LSA.

De même, Bosch et ses collègues [Bosch et al., 2006] calculent des sacs-de-mots à partir des descripteurs locaux SIFT, déterminés à partir d'une grille régulière. La représentation en sacs-de-mots des images est utilisée pour l'apprentissage des objets ou catégories de l'image, à l'aide du modèle pLSA. Puis un classifieur des plus proches voisins est appliqué sur la distribution des objets dans les images afin de les classifier. Les performances obtenues par cette approche sont très satisfaisantes.

Par ailleurs, Blei et Jordan [Blei & Jordan, 2003] utilisent des paires de modules LDA pour modéliser les relations par exemple entre les images et les légendes qui leur correspondent, mais aussi entre les articles et leur bibliographie, etc. Chaque image d'apprentissage annotée avec 2 à 4 mots, est segmentée en une dizaine de régions maximum. Et le vocabulaire de 168 mots est obtenu à partir de caractéristiques extraites des régions de l'image, puis quantifiées. Les auteurs testent trois modèles hiérarchiques de génération de données annotées : le *Gaussian-multinomial mixture model* (GM-mixture) dans lequel les régions et les mots sont générés indépendamment et la correspondance entre les régions et les mots est ignorée, le *Gaussian Multinomial LDA* (GM-LDA) qui capture la possibilité d'avoir des mots et des régions issues du même modèle, mais ne fait pas de lien direct entre la description d'une région et son annotation, et le *correspondence LDA* (corr-LDA) qui combine la flexibilité de GM-LDA et l'associativité de GM-mixture. Ils montrent que ce dernier modèle a de meilleures performances et annote correctement la plupart des images de test, comme illustré dans la figure 5.4.

Toutes ces approches d'annotation ont chacune leurs avantages et leurs inconvénients, et le choix d'un type d'approche est fonction des données, des moyens dont on dispose, et de l'application visée. Par ailleurs, elles ont toujours été testées sur des images multimédia. Ici, nous nous questionnons sur l'efficacité de telles méthodes sur les images satellitaires, et nous intéressons aux modèles génératifs probabilistes à variable latente de l'analyse textuelle, qui ont été adapté avec succès pour le traitement des images multimédia.



FIG. 5.4 – Exemples d’images de l’ensemble de test et leurs annotations automatiques obtenus avec les différents modèles présentés dans [Blei & Jordan, 2003], et basés sur une représentation en sacs-de-mots.

5.4.2 Modèles de mélange de l’analyse de texte

Comme indiqué dans la section 5.4.1.2, des modèles génératifs initialement développés pour la modélisation de grands corpus textuels, ont récemment été utilisés avec succès pour la classification et l’annotation des images multimédia. Dans cette section, nous nous intéressons à de tels modèles probabilistes, plus particulièrement à l’analyse sémantique latente probabiliste (*probabilistic Latent Semantic Analysis* ou pLSA) et à l’Allocation Dirichlet Latente (*Latent Dirichlet Allocation* ou LDA). Ces deux modèles sont basés sur l’hypothèse de “sacs-de-mots”, qui suppose que l’ordre des mots dans un document peut être négligé. Cette hypothèse, dite d’échangeabilité [Aldous, 1985], est aussi valable pour les documents d’un corpus. Par ailleurs, un théorème, mis en évidence par De Finetti (1990), établit que toute collection de variables aléatoires échangeables a une représentation sous la forme d’un mélange de distributions. Ainsi, les modèles pLSA et LDA sont des modèles de mélange. Par rapport au modèle unigramme, un modèle de mélange tient compte de l’hétérogénéité des données, et permet ainsi de mieux les décrire, moyennant une complication de modèle. Dans le cas des modèles pLSA et LDA, les poids des différents modèles dans le mélange représentent les poids des “topics” latents, dont la structure capture la sémantique du document.

5.4.2.1 Analyse sémantique latente probabiliste (pLSA)

Proposé par Hofmann [Hofmann, 1999], le modèle pLSA est un modèle génératif très utilisé dans le domaine de la modélisation de textes. C’est l’une des premières méthodes à fournir une approche probabiliste vers la modélisation des documents textuels comme des mélanges de topics latents. En fait, le modèle pLSA suppose l’existence d’une va-

riable latente z (topic) dans le processus génératif de chaque mot w_n d'un document \mathbf{w} . Etant donnée cette variable cachée, chaque mot w_n est indépendant du document à partir duquel il a été généré. La probabilité jointe des variables observées est donc :

$$p(\mathbf{w}, w_n) = p(\mathbf{w}) \sum_z p(w_n | z)p(z | \mathbf{w}) \quad (5.2)$$

Le modèle pLSA capture ainsi la possibilité qu'un document contienne plusieurs topics, $p(z | \mathbf{w})$ représentant la proportion de chaque topic dans le document \mathbf{w} . Cependant, le modèle pLSA n'apprend les probabilités $p(z | \mathbf{w})$ que pour les documents de l'ensemble d'apprentissage. Par conséquent, il n'est pas un modèle génératif bien défini, car il n'existe aucun moyen direct d'attribuer une probabilité à un document n'appartenant pas à l'ensemble d'apprentissage. En outre, le nombre de paramètres à estimer croît linéairement avec la taille de l'ensemble d'apprentissage, ce qui peut provoquer des problèmes de surapprentissage. Le modèle LDA résout ces limitations en considérant que les poids des topics sont générés à partir d'une distribution de Dirichlet commune, plutôt que de les considérer comme un grand ensemble de paramètres individuels explicitement liés à l'ensemble d'apprentissage. Pour ces raisons, le modèle LDA est très souvent préféré au modèle pLSA.

5.4.2.2 Allocation Dirichlet Latente (LDA)

Le modèle LDA [Blei et al., 2003b] est un modèle génératif probabiliste pour les collections de données discrètes telles que les corpus textuels. Ce modèle bayésien hiérarchique à trois niveaux représente les documents d'un corpus comme des mélanges aléatoires sur des topics latents obtenus de manière non supervisée. Chaque topic est à son tour, caractérisé par une distribution sur des mots. Le modèle LDA suppose le processus génératif suivant (algorithme 5.4.2.2), pour chaque document \mathbf{w} appartenant à un corpus D :

Algorithme 5.4.2.2 : Processus génératif du modèle LDA

- 1 : Tirer N , suivant une loi de Poisson de paramètre ξ
 - 2 : Tirer θ suivant une distribution de Dirichlet de paramètre α , θ étant de dimension K .
 - 3 : Pour chaque position n des N mots du document :
 - 3.1 : Tirer un topic z_n selon une loi multinomiale de paramètre θ ,
 - 3.2 : Tirer un mot w_n suivant la loi $p(w_n | z_n, \beta)$, loi multinomiale conditionnée par le topic z_n .
-

où :

- N est le nombre de mots dans le document $\mathbf{w} = (w_1, \dots, w_N)$. N étant indépendant des autres données générant les variables, le fait qu'elle soit une variable aléatoire est en général ignoré.
 - La dimensionnalité K de la variable aléatoire θ , et donc celle de la variable topic z , est supposée connue. θ peut prendre ses valeurs dans le $(K - 1)$ simplexe (car $\theta_k \geq 0$ et $\sum_{k=1}^K \theta_k = 1$), et chaque θ_k indique la proportion du topic k dans le document \mathbf{w} .
-

- V étant la taille du vocabulaire, β est une matrice $K \times V$ de paramètres à estimer telle que $\forall k \in \{1 \dots K\}$, $\sum_{v=1}^V \beta_{kv} = 1$. Les β_{kv} sont les probabilités des mots sachant le topic : $\beta_{kv} = p(w = v | z = k)$.

Ainsi, la probabilité de générer un document \mathbf{w} sachant les paramètres α et β est :

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta \quad (5.3)$$

L'ordre des documents dans le corpus étant également supposé non pertinent, la vraisemblance du corpus D est donc déterminée par le produit des probabilités marginales des différents documents :

$$p(D | \alpha, \beta) = \prod_{d=1}^M p(\mathbf{w}_d | \alpha, \beta) \quad (5.4)$$

Cette grandeur est utilisée durant la phase d'estimation du modèle LDA, qui consiste à déterminer les paramètres α et β qui maximisent la log-vraisemblance de la collection de documents. Comme nous l'avons mentionné plus haut, si le modèle LDA permet de mieux décrire les données, l'estimation de ses paramètres est en revanche, plus ardue. En effet, l'équation 5.3 n'étant pas traitable pour une inférence exacte, la solution consiste à utiliser des algorithmes d'approximation tels que la méthode *Expectation - propagation* [Minka & Lafferty, 2002], l'échantillonneur de Gibbs [Griffiths & Steyvers, 2002] ou encore l'inférence variationnelle [Blei et al., 2003b]. Cette dernière méthode, qui selon Blei et ses collègues permet d'estimer correctement les paramètres, exploite l'inégalité de Jensen pour obtenir une borne inférieure de la log-vraisemblance. Cette borne peut être maximisée par rapport à α et β , via une procédure de maximisation de l'espérance (EM) avec des paramètres variationnels.

Une fois les paramètres du modèle déterminés, il est alors possible d'attribuer une probabilité à un document n'appartenant pas à l'ensemble d'apprentissage. Cette propriété fait du LDA, un modèle génératif complet.

Les avantages des modèles génératifs tels que le LDA incluent leur modularité et leur extensibilité. En effet, le LDA peut être utilisé comme un module d'un modèle plus complexe. Dans de récents travaux, Blei et Jordan [Blei & Jordan, 2003] utilisent des paires de modules LDA pour modéliser les relations entre deux types de données, l'un étant l'annotation de l'autre (par exemple des articles et leur bibliographie). En outre, plusieurs modèles, dérivés du LDA ont été proposés dans la littérature. On peut citer, entre autres, le *Correlated topic model* [Blei & Lafferty, 2006] qui expose la corrélation entre les proportions de topics, en utilisant la distribution normale logistique plutôt que la distribution de Dirichlet, et le *hierarchical LDA* [Blei et al., 2003a] dans lequel les topics sont regroupés dans une hiérarchie.

Dans un autre registre, se trouve le modèle *Spatial LDA* (SLDA) [Wang & Grimson, 2008], qui s'attaque au problème de l'information spatiale, si importante dans plusieurs travaux de vision par ordinateur, mais pourtant négligée dans le modèle LDA. En effet, le modèle SLDA encode la structure spatiale des mots visuels, cette information spatiale étant utilisée dans la conception des documents. Contrairement au modèle LDA dans lequel le partitionnement des mots dans le document est connu a priori, l'attribution mots / documents dans le SLDA est une variable aléatoire cachée. Il existe un processus

génératif, dans lequel la connaissance de la structure spatiale peut être rajoutée comme étant un a priori, groupant ainsi les mots visuels proches dans l'espace dans le même document. Dans le chapitre 7, nous considérons ce problème de l'information spatiale dans l'utilisation du modèle LDA pour l'annotation sémantique de grandes images satellites.

5.4.2.3 Utilisation pour la modélisation des images

Tel que mentionné dans la section 5.4.1.2, les modèles statistiques de fouille de textes comme la pLSA et le LDA ont été utilisés pour modéliser les images. Cependant, l'utilisation de ces modèles implique que les entités soient représentées sous une forme particulière, dite en sacs-de-mots, qui suppose la non pertinence de l'ordre d'apparition des mots et l'existence d'un vocabulaire fixe et fini. En analyse de texte, cela se traduit par un vocabulaire qui peut se limiter aux radicaux des mots. Cependant, comme indiqué précédemment, la notion de "mot" est moins évidente en ce qui concerne les images, qui par conséquent, n'ont pas de vocabulaire a priori. Afin de construire un vocabulaire pour la collection d'images, il est nécessaire tout d'abord, d'établir une équivalence entre les différentes terminologies, c'est-à-dire, de définir à quoi correspondent un mot, un document et un corpus pour les images.

Mots et documents pour les images Dans le domaine textuel, chacun des mots du vocabulaire exprime une idée, et apporte indépendamment des autres, une information importante pour le document. Cette information à elle-seule n'est pas forcément discriminante pour identifier le document, cependant, elle donne des indications sur sa nature. Un document étant formé de plusieurs mots du vocabulaire dans des proportions différentes, les informations apportées par les différents mots sont combinées pour caractériser le document. Dans le domaine des images, et plus particulièrement en imagerie satellitaire, le choix d'un mot dépend de taille de l'image, de sa résolution, de l'objectif visé, ainsi que du niveau de sémantique des concepts identifiés dans l'image par un utilisateur. Du choix du type de mot, dépendent le vocabulaire et le type du document. En effet, si l'on souhaite par exemple classifier les régions d'une grande image, un mot peut être un pixel de l'image, et dans ce cas, un document est un segment de l'image. Le corpus peut ainsi correspondre à l'ensemble des segments de la grande image. Par ailleurs, pour des opérations de recherche d'images par le contenu dans de grandes bases, un mot peut être une fenêtre de pixels ou une région de l'image obtenue via une segmentation. Le document correspond alors à l'image elle-même, et le corpus est l'ensemble des images de la base. Par exemple, les mots utilisés par Fei-Fei et Perona [Fei-Fei & Perona, 2005], dans leurs travaux de reconnaissance de la sémantique des scènes naturelles, sont des fenêtres de pixels (figure 5.2). Cependant, dans la littérature, la plupart des travaux représentent les mots par les segments de l'image, qui ont l'avantage de partitionner l'image de manière plus naturelle, plutôt que par les fenêtres de pixels obtenues par découpage suivant une grille [Barnard et al., 2003; Monay & Gatica-Perez, 2003].

Construction du vocabulaire de mots visuels Une fois que les mots sont définis, l'étape suivante est de construire le vocabulaire visuel, comme détaillé dans la section 5.2, afin de réduire le nombre de mots dans la base d'images. Malgré tout, nous en présentons un résumé succinct pour clarifier la notion de "mot visuel", à ne pas confondre avec

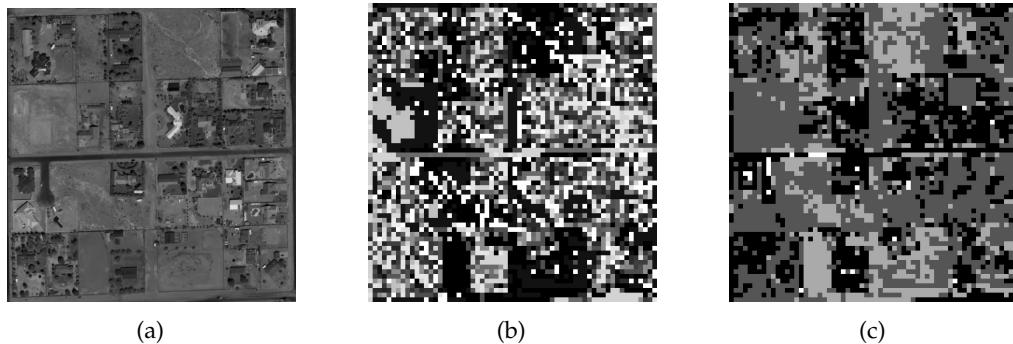


FIG. 5.5 – Image de *banlieue résidentielle* issue d’une image satellitaire Quickbird de Las Vegas à 60 cm de résolution (a), ainsi que sa représentation à l’aide des mots visuels (b), et des topics obtenus de manière non supervisée (c).

“mot”. Des caractéristiques de bas-niveau (radiométrie, texture, etc) sont automatiquement extraites de chaque mot de la base d’images. Une quantification vectorielle est ensuite appliquée sur l’ensemble des primitives de manière à les répartir dans des clusters, nous permettant ainsi de travailler à partir d’une collection discrète de caractéristiques. En effet, deux mots appartenant au même cluster sont considérés équivalents et sont représentés par le vecteur quantifié : le mot visuel. Le nombre de mots visuels V est la taille du vocabulaire. Etant donné le vocabulaire, chaque image I est représentée par une séquence de N mots visuels. Le compte des occurrences des mots visuels engendrant un vecteur de fréquences pour chaque image de la collection, la base d’images peut alors être décrite dans une matrice de cooccurrence de taille $V \times M$ où M est le nombre d’images dans la base.

La figure 5.5(a) montre une image de *banlieue résidentielle*, issue d’une image satellitaire Quickbird de Las Vegas à 60 cm de résolution, et la figure 5.5(b) montre la même image dans laquelle chaque mot est remplacé par le mot visuel correspondant. Les mots visuels sont obtenus à partir de fenêtres de taille 10×10 pixels, issues du découpage de l’image suivant une grille régulière, puis quantifiées.

Détermination des topics Le modèle LDA modélise chaque mot d’une image comme un échantillon d’un modèle de mélange, où les composantes du mélange peuvent être vues comme des représentations des topics. Il est ainsi possible d’assigner un topic différent à chaque mot de l’image, et plusieurs mots peuvent appartenir au même topic latent. Cette structure latente, obtenue de manière non supervisée, peut capturer une information sémantique dans l’image. Les figures 5.5(c) et 5.6 montrent un exemple de la distribution mots / topics respectivement dans l’image et dans l’espace des primitives, obtenue à partir de l’image de *banlieue résidentielle* de la figure 5.5(a). Les principaux topics qui y apparaissent ont pu être identifiés comme étant les maisons et arbres (bleu), les terrains nus (vert), les terrains de sable (rouge). Le quatrième topic (noir) et le dernier (rose), très peu représentés, ne correspondent à aucun objet en particulier. Notons en outre que des mots éloignés dans l’espace des attributs peuvent être regroupés dans le même topic (par exemple, le topic rouge). Cette manière de regrouper des mots ayant des caractéristiques physiques différentes est un pas vers l’émergence de la sémantique, et peut au moins donner des indications sémantiques fortes. Ainsi, un topic peut correspondre à un type

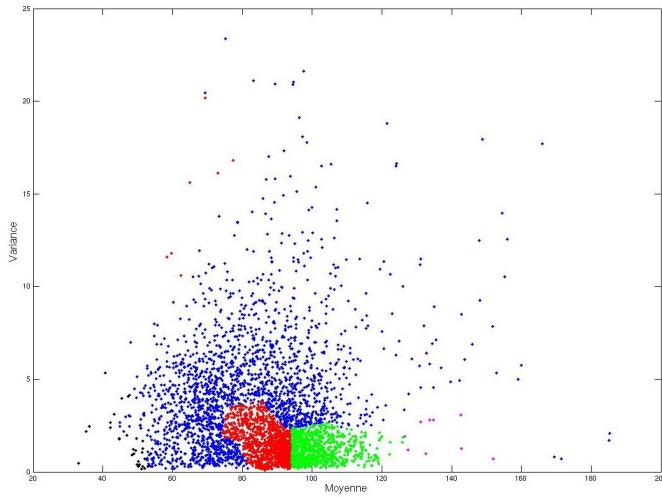


FIG. 5.6 – Distribution mots / topics dans l'espace des primitives (moyenne et variance), de l'image de la figure 5.5(a). Chaque couleur correspond à un topic, et nous pouvons attribuer une sémantique aux principaux topics - bleu : maisons et arbres, vert : terrains nus, rouge : terrains de sable, noir et rose : aucun objet en particulier.

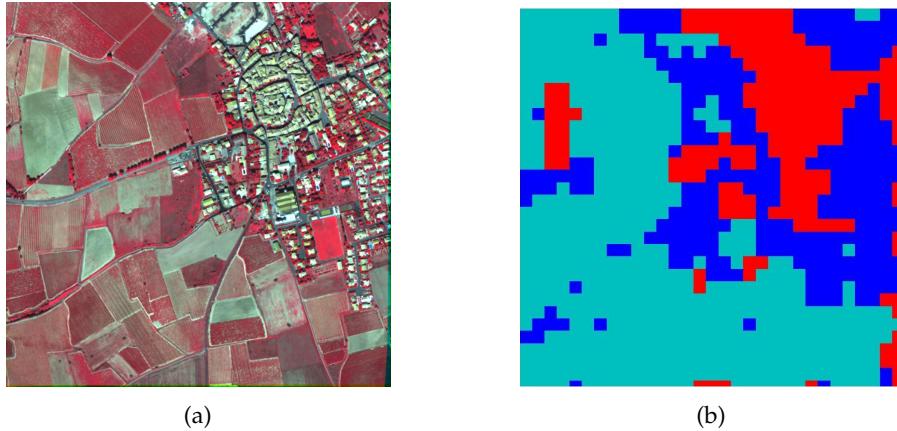


FIG. 5.7 – Exemple d'une image multispectrale de Roujan ©CNES, à 1 m de résolution (a), et la distribution des topics dans l'image (b). Chaque couleur représente un topic, qui correspond ici à une partie de l'image : les bâtiments, région périurbaine et champs cultivés. Les caractéristiques utilisées sont radiométriques.

d'objet ou de partie d'objet de l'image, et le mélange de topics latents indique les proportions de chaque type d'objet dans l'image.

Précisons que dans cet exemple, nous n'avons utilisé que la moyenne et la variance comme primitives pour la construction du vocabulaire. L'utilisation de primitives plus riches peut aboutir à la détermination de topics se rapprochant de plus en plus des parties ou des objets de l'image (figure 5.7). Et bien sûr, le choix du nombre de mots visuels et du nombre de topics est déterminant pour faire apparaître la sémantique. Nous considérons ce problème dans le chapitre 7.

5.5 Conclusions

Dans ce chapitre, nous avons introduit et justifié notre choix de représentation des images, motivé par les difficultés rencontrées lors d'une étude précédente pour l'identification des classes de mélange. En effet, nous avons analysé le potentiel de la représentation de l'image par des mots visuels, et montré son intérêt pour notre problème, en l'occurrence, elle permet de tirer profit des techniques textuelles. En outre, nous avons précisé dans les détails, les différentes étapes de construction de notre vocabulaire visuel. Les différentes directions, pour lesquelles nous avons adopté cette représentation, c'est-à-dire l'exploitation du principe de compositionnalité sémantique et l'utilisation des modèles probabilistes de l'analyse de textes basés sur une représentation en sacs-de-mots, ont été présentées, ainsi que les réflexions sur leur utilisation pour les images. Ces applications sont explorées dans les chapitres suivants. En effet, le chapitre 6 étudie les relations spatiales d'adjacence entre les mots visuels afin de déterminer des classes mélangées, tandis que le chapitre 7 s'intéresse à un modèle textuel : le LDA, qui, au terme de l'étude approfondie que nous avons faite dans ce chapitre, nous semble pertinent pour notre tâche d'association des régions de l'image à des classes sémantiques simples et de mélange.

Chapitre 6

Relations spatiales entre les mots visuels

Dans ce chapitre, nous nous proposons d'identifier les classes de mélange dans les images satellitaires en exploitant le principe de compositionnalité sémantique, défini dans le chapitre 5. Ce principe stipule que la signification d'une expression complexe est fonction de la signification de ses parties et de leurs relations syntaxiques. En ce qui concerne les images, après la détermination des mots visuels, décrits dans le chapitre 5, il s'agit de tirer profit des relations spatiales, plus particulièrement des relations d'adjacence entre ces mots visuels.

Dans un premier temps, nous proposons une approche non supervisée, pour déterminer les structures et objets d'intérêt de l'image. Ces derniers, obtenus par regroupement des mots visuels suivant un critère à définir, peuvent éventuellement correspondre à une classe sémantique.

Nous décrivons ensuite une approche supervisée d'apprentissage des relations d'adjacence entre les mots visuels d'une classe sémantique donnée (une classe de mélange), et sa généralisation sur des zones inconnues.

Cependant, l'exploitation des relations d'adjacence entre des mots visuels voisins dépend fortement du mode de représentation de l'information sous-jacente. Dans la section suivante, nous proposons donc d'abord une analyse, puis une discussion des différents choix possibles de représentation.

6.1 Représentation des informations extraites

Le mode de représentation des informations extraites de l'image conditionne la manière dont l'image, une fois codée (décomposée en mots visuels), est manipulée. Le choix d'une bonne représentation est important car de cette représentation, dépend l'aisance de l'utilisateur à exploiter ces informations.

Dans notre étude, nous nous sommes intéressés à trois types de représentation des connaissances : les représentations par graphes plus utilisées dans la littérature, les représentations par hypergraphes qui sont une généralisation des graphes, et enfin les représentations de l'image par un texte ou une chaîne.

6.1.1 Les graphes d'adjacence

Un grand nombre de systèmes ou structures peuvent être considérés de façon abstraite comme un ensemble d'éléments, certains étant en relation, d'autres pas. Il est alors souvent profitable de les représenter à l'aide d'un graphe [Fournier, 2007]. En effet, ils proposent une représentation compacte, structurée, complète et facile à manipuler.

Les graphes d'adjacence décrivent les relations de voisinage entre les éléments de l'image. L'image peut donc être représentée par un graphe d'adjacence connexe, dans lequel les pixels (graphe de pixels) ou les régions élémentaires (graphe d'adjacence des régions) constituent les sommets du graphe. Classiquement, il existe, aux variantes près, trois façons de représenter les graphes en vue d'un traitement algorithmique : la matrice d'adjacence, les listes d'adjacence et la liste des arêtes. La représentation par matrice d'adjacence est simple et commode par exemple, pour tester l'existence d'une arête entre deux sommets donnés. Son principal inconvénient est son encombrement mémoire qui est en $O(|X|^2)$ où $|X|$ est le nombre de sommets du graphe. L'idée de la représentation par listes d'adjacence est de s'affranchir du format imposé d'une matrice et de ne stocker que les informations pertinentes, correspondant aux arêtes. Ainsi, elle est plus adaptée pour représenter les graphes creux. La représentation par listes d'arêtes, quant à elle, est semblable à celle par listes d'adjacence, sauf que le mode d'entrée dans le graphe se fait par les arêtes.

Les utilisations de la théorie des graphes sont variées, avec des applications dans les domaines du traitement d'images (segmentation, détection de contours), de la reconnaissance des formes ou encore l'interprétation des scènes structurées. Ce succès est en partie dû au fait qu'il existe plusieurs algorithmes pour extraire, manipuler et exploiter l'information contenue dans les graphes : entre autres, les parcours de graphes (parcours en largeur d'abord et en profondeur d'abord), la recherche de l'arbre recouvrant de poids minimal, le flot maximum (et coupe de capacité minimale) et les recherches de chemins optimaux.

6.1.2 Les hypergraphes

Malgré leur efficacité, les graphes présentent un inconvénient majeur résidant dans le fait que les relations entre les sommets sont binaires. En effet, ce type de relations ne peut pas bien traduire le processus de structuration d'objets de nature arbitraire. Tenant compte de ce handicap, Berge a généralisé le concept de graphe en introduisant les *hypergraphes* [Berge, 1973], où les relations de voisinage ne sont plus exclusivement binaires.

Définition : Soit une image I représentée par :

$$I : X \subseteq \mathbb{Z}^2 \rightarrow \Omega \subseteq \mathbb{Z}^n \quad \text{avec } n \geq 1 \quad (6.1)$$

où les éléments de X sont les coordonnées des pixels de l'image et Ω est l'ensemble de leurs caractéristiques radiométriques. Un hypergraphe H sur un ensemble X est une famille $(E_i)_{i \in I}$ de parties non vides de X appelées *hyperarêtes*, dont la réunion sur I est X . Mathématiquement :

$$\begin{aligned} \forall i \in I, E_i &\neq \emptyset \\ \bigcup_{i \in I} E_i &= X. \end{aligned} \quad (6.2)$$

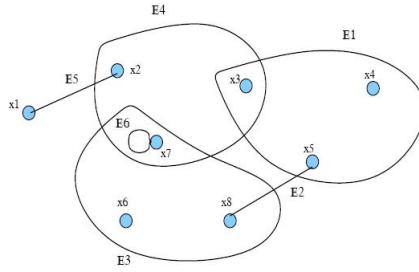


FIG. 6.1 – Représentation d'un hypergraphe.

Les éléments de X sont, comme dans la théorie des graphes, appelés *sommets*, alors que les éléments $(E_i)_{i \in I}$ sont des *hyperarêtes*. Ainsi, la notion de graphe est bien généralisée par celle d'hypergraphe puisque les arêtes y sont remplacées par des ensembles de taille arbitraire.

Un hypergraphe H est souvent représenté en dessinant sur le plan, des points représentant ses sommets. L'hyperarête E_j sera représentée par un trait continu joignant ses deux éléments si $|E_j| = 2$, par une boucle si $|E_j| = 1$, et par un trait plein entourant ses éléments si $|E_j| \geq 3$ (voir figure 6.1).

Hypergraphe de voisinage d'une image à niveau de gris : Soient une distance colorimétrique d' sur Ω et une distance de grille d sur X . Soient α et β , deux réels strictement positifs. A chaque pixel x de X , il est possible d'associer un unique voisinage $\Gamma_{\alpha,\beta}(x)$ pour l'image I :

$$\Gamma_{\alpha,\beta}(x) = \{y \in X, y \neq x \text{ tel que } d'(I(x), I(y)) \leq \alpha \text{ et } d(x, y) \leq \beta\} \quad (6.3)$$

α et β sont appelés respectivement seuil *colorimétrique* et seuil *spatial*. $\Gamma_{\alpha,\beta}$ est le *voisinage spatiocolorimétrique*, qui permet de décrire un certain niveau de cohérence (homogénéité) entre un pixel et son environnement dans l'image, ceci sous réserve que les seuils α et β soient suffisamment "bien" choisis.

Ainsi, à partir de chaque pixel $x \in X$ et de $\Gamma_{\alpha,\beta}(x)$, il est possible d'associer pour chaque image I , un hypergraphe de voisinage $H_{\alpha,\beta}$ défini comme :

$$H = (X, E_{\alpha,\beta}(x)), \text{ avec } E_{\alpha,\beta}(x) = (\{x\} \cup \Gamma_{\alpha,\beta}(x))_{x \in X} \quad (6.4)$$

$E_{\alpha,\beta}(x)$ étant l'hyperarête centrée en x .

Il est évidemment possible d'associer à tout graphe (X, E) , un hypergraphe de la manière suivante : à chaque sommet x du graphe, on peut associer le voisinage $\Gamma(x)$ de ce sommet, défini comme l'ensemble :

$$\Gamma(x) = \{y \in X \text{ tel que } \{x, y\} \in E\} \quad (6.5)$$

Ainsi, l'ensemble $\{x\} \cup \Gamma(x)$ est l'hyperarête engendrée par le sommet x du graphe.

En ce qui concerne les images en couleur, l'application de la représentation par hypergraphes de voisinage peut se faire suivant deux approches : l'approche marginale dans laquelle chacune des composantes est traitée séparément, et l'approche vectorielle où on utilise des vecteurs plutôt que des intensités scalaires dans une image à niveau de gris.

6.1.3 Représentation de l'image en un texte

Cette méthode, présentée dans [Watanabe et al., 2002], a déjà été évoquée dans le chapitre précédent pour exposer un cas particulier de codage de l'image en mots visuels. Ici, nous insistons sur la technique de prise en compte des relations spatiales entre ces mots visuels qui y est proposée. Il s'agit de transformer l'image en un texte, ce qui présente l'avantage de la rendre linéaire.

Pour générer le texte ou la chaîne, on construit d'abord le graphe pondéré des pixels de l'image, dans lequel le poids d'une arête est la différence spectrale (calculée en utilisant une fonction distance appropriée) entre ses extrémités. Puis, l'arbre recouvrant de poids minimal est construit à partir de ce graphe, en supprimant progressivement les arêtes de manière à ce que la somme des poids soit minimale. Partant d'un noeud initial (en général en haut à gauche de l'image), l'arbre est parcouru suivant les arêtes de poids les plus faibles. Ce qui produit au final une séquence qui, pour chaque pixel parcouru, donne un couple formé du vecteur d'attributs du pixel, et la direction traversée du noeud précédent au noeud courant dans l'arbre. L'ensemble des directions possibles peut être défini suivant le code de Freeman. Enfin, une quantification de la séquence permet d'obtenir un texte code pour l'image, pour lequel la taille de l'alphabet est le nombre de groupes obtenus. En effet, chaque pixel est associé à une lettre de l'alphabet, qui est fonction des caractéristiques du pixel et du noeud voisin le précédent lors du parcours de l'arbre. Watanabe et ses collègues utilisent cette représentation de l'image pour générer des taux de compression entre des images de référence et une image test. Cette dernière sera classée avec l'image de référence avec qui elle a le plus petit taux de compression.

Le plus intéressant pour nous est qu'il est également possible de travailler sur des régions au lieu des pixels. Dans ce cas, l'image est d'abord segmentée et l'arbre recouvrant de poids minimal est extrait du graphe pondéré d'adjacence des régions. Au final, l'image est représenté par un texte dans lequel chaque région a un identifiant qui dépend des caractéristiques de la région, mais aussi de la région voisine l'ayant précédé dans le balayage de l'arbre.

6.1.4 Analyse et choix d'une représentation

Rappelons que nous souhaitons déterminer les structures sémantiques de l'image, en exploitant les relations spatiales et contextuelles entre les mots visuels. Nous avons donc testé les différents modes de représentation sus-présentés, afin de choisir la mieux appropriée pour notre objectif.

En ce qui concerne les hypergraphes, l'idée est de mêler segmentation et reconnaissance pour mettre directement en évidence les structures de l'image. Pour ce faire, nous nous proposons d'exploiter les outils de partitionnement des hypergraphes pour la segmentation d'images, les relations spatiales et contextuelles étant alors prises en compte directement dans la segmentation. La procédure que nous avons adoptée est la suivante. A partir de l'image multispectrale, l'hypergraphe de voisinage spatiocolorimétrique de l'image est construit. Nous avons choisi de suivre une approche vectorielle qui prend en compte les dépendances entre les composantes. La distance colorimétrique permettant de générer les hyperarêtes est la distance euclidienne, appliquée dans une grille rectangulaire et un 8-voisinage.

La figure 6.2(a) montre une image dont l'hypergraphe a été construit en utilisant un seuil colorimétrique fixé à 35. Afin de visualiser le comportement de l'hypergraphe ob-

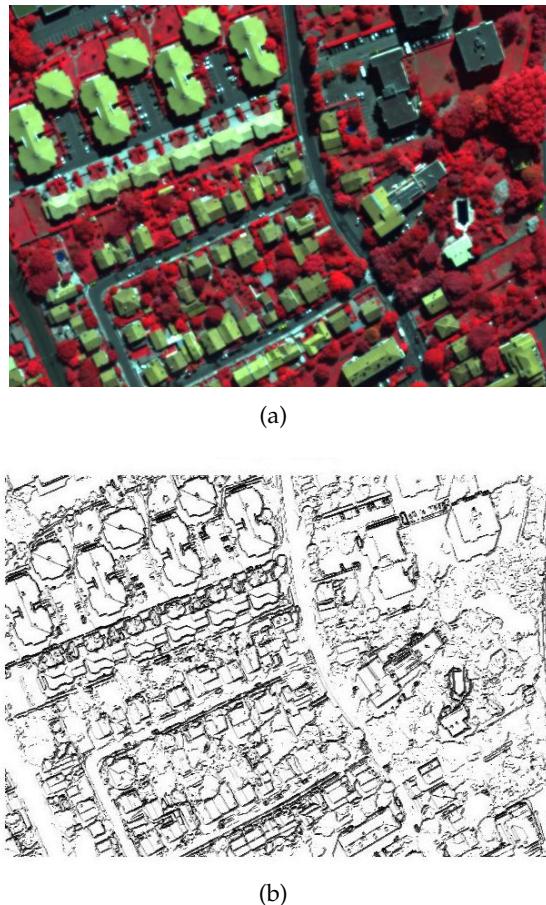


FIG. 6.2 – Visualisation du comportement d'un hypergraphe de voisinage spatiocolorimétrique vectoriel avec un seuil de 35. (a) Image Pelican de Toulouse, à 1 m ; (b) Chaque pixel de l'image a une intensité qui est fonction du nombre de pixels de son voisinage qui appartiennent à l'hyperarête dont il est le centre : les zones blanches sont les plus homogènes.

tenu, nous présentons dans la figure 6.2(b), la même image dans laquelle chaque pixel a un niveau de gris dépendant du nombre de pixels de son voisinage qui appartiennent à l'hyperarête centrée en ce pixel : plus un pixel a de voisins appartenant à l'hyperarête dont il est le centre, plus son intensité est forte. Donc dans la figure 6.2(b), les zones blanches sont les plus homogènes.

Le but d'un k-partitionnement d'hypergraphes¹ est de partitionner les noeuds de l'hypergraphe en k sous-ensembles disjoints, de telle sorte qu'une certaine fonction définie sur les hyperarêtes soit optimisée (minimisation de la coupe des hyperarêtes, minimisation de la somme des degrés externes). Une représentation par hypergraphes paraît donc intéressante pour représenter plus efficacement le contenu de l'image de façon à faire apparaître les structures d'intérêt pour l'utilisateur. Mais cela nécessite de définir une fonction adéquate sur les hyperarêtes (par exemple une fonction de saillance particulière), ce qui n'est guère trivial. Et lorsqu'elle est mal choisie, cela peut aboutir à un

¹Pour ce faire, il existe un outil de partitionnement d'hypergraphes appelé *HMETIS* : <http://glaros.dtc.umn.edu/gkhome/metis/hmetis/overview>

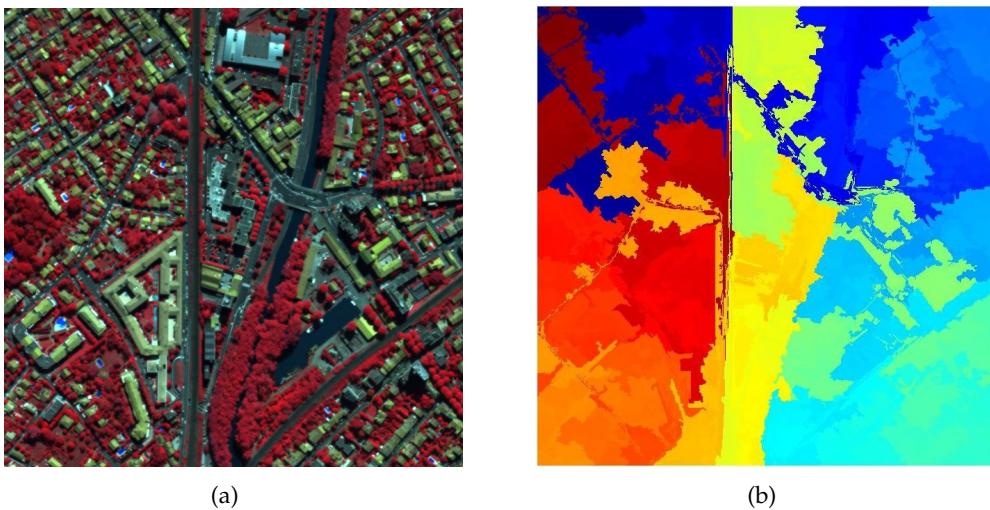


FIG. 6.3 – Test de codage de l'image en un texte, puis segmentation de l'image par partitionnement du texte. (a) Image Pelican de Toulouse à 1 m de résolution ; (b) Image segmentée : une région est une suite de lettres dans le texte et les frontières des régions sont données par les noeuds terminaux dans l'arbre recouvrant de poids minimal. Les couleurs sont distribuées de manière aléatoire.

partitionnement de l'image totalement non intuitif.

Par ailleurs, nous avons testé la représentation de l'image en un texte sur une image Pelican. L'idée était de transformer l'image en un texte et d'utiliser cette représentation linéaire qui conserve l'information spatiale (puisque la suite de lettres est obtenue par parcours de l'arbre recouvrant de poids minimal) pour identifier des suites de lettres ou des sous-séquences particulières pouvant représenter des objets ou structures particuliers de l'image. Un partitionnement judicieux (suivant un critère d'entropie minimale par exemple) du texte code de l'image pourrait nous aider à parvenir à nos fins. De même, l'utilisation de méthodes a contrario, afin de mettre en évidence des sous-séquences singulières de la chaîne, pourrait être une alternative intéressante. Cependant, les difficultés liées à la manipulation et à l'exploitation des informations contenues dans le texte pour nos besoins, ainsi que la longueur du texte pour de grandes images se sont avérées très contraignantes. En effet, les tests de partitionnements du texte que nous avons effectués ne nous ont guère donné de résultats convaincants. La figure 6.3 montre un exemple de segmentation de l'image par partitionnement du texte, où les frontières des régions sont données par les noeuds terminaux des brins de l'arbre pendant le parcours selon 4 directions. Outre les structures linéaires qui sont assez bien détectées, les autres structures ou objets de l'image sont assez confuses. Différents autres tests ont été effectués sur des images synthétiques plus simples, généralement avec des résultats décevants.

Précisons que si dans ces expérimentations nous avons travaillé avec le pixel, cela est tout à fait généralisable au niveau de la région, en segmentant l'image au préalable et en considérant un noeud comme étant une région. Par ailleurs, ces deux modes de représentation, d'une manière ou d'une autre, sont basés sur les graphes qui sont un outil classique pour ce type de problème. Donc dans la suite, nous utiliserons les graphes

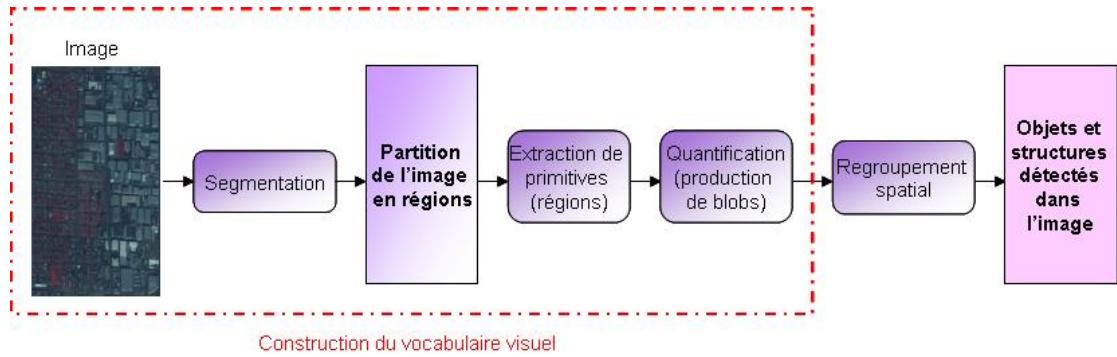


FIG. 6.4 – Chaîne de traitement définie pour la détection des structures dans une image.

d'adjacence comme mode de représentation dans nos expérimentations. Nous en ferons usage comme d'un outil pour le regroupement spatial des régions.

6.2 Détermination non supervisée de structures d'intérêt

Suivant le principe de compositionnalité sémantique, l'image est d'abord décomposée en zones élémentaires à l'aide d'une segmentation. Notons qu'il est également possible de détecter des formes particulières de l'image telles que les réseaux routiers ou fluviaux ; cependant, nous n'en tiendrons pas compte dans cette expérimentation. Les zones élémentaires sont ensuite quantifiées pour créer les mots visuels. Puis, de même que la compréhension d'une phrase nécessite de comprendre les mots et la syntaxe entre ces mots, nous exploitons les relations existant entre ces mots visuels, en particulier les relations d'adjacence. Les mots visuels peuvent donc être regroupés suivant leurs attributs géométriques et/ou topologiques afin de former des structures d'intérêt, c'est-à-dire saillantes au regard d'un critère à définir.

La figure 6.4 récapitule les étapes que nous avons définies pour la mise en évidence des structures potentiellement intéressantes de l'image.

6.2.1 Protocole

La procédure suivie est celle schématisée dans la figure 6.4. L'image est donc d'abord partitionnée en régions élémentaires. Après l'étape de quantification, une étiquette est associée à chaque cluster, représenté par un mot visuel. Chaque région de l'image est ainsi doté d'une étiquette, signifiant son appartenance à un cluster. Par la suite, nous ne manipulons plus les caractéristiques de bas niveau et nous intéressons aux relations d'adjacence entre les régions.

Pour représenter cette adjacence, nous avons utilisé les graphes. Ainsi, nous avons créé le graphe pondéré d'adjacence des régions, dans lequel le poids d'une arête détermine le degré d'adjacence entre ses deux extrémités. Le degré d'adjacence représente la force du lien entre les deux régions voisines. Il peut s'exprimer de plusieurs manières :

- la compacité ou la forme de la région finale obtenue par regroupement, ceci afin de juger de la régularité de la forme de la région obtenue. elle peut être déterminée à l'aide du rapport isopérimétrique que nous détaillons dans la section suivante.

- la longueur relative de la frontière commune, c'est-à-dire le nombre de pixels appartenant aux deux régions adjacentes normalisé par le périmètre de la région finale.
- la distance entre les centres de gravité des deux régions élémentaires.
- ...

Des opérations sur les graphes sont ensuite menées, pour détecter de manière non supervisée, les objets et structures de l'image. L'algorithme proposé est le suivant (algorithme 6.2.1) :

Algorithme 6.2.1 : Pseudo-code pour la détection non supervisée des structures.

- 1 : Construire le graphe pondéré d'adjacence des régions,
 - 2 : Calculer une matrice de probabilités $P_{i,j}$, représentant la probabilité qu'une région d'étiquette i soit adjacente à une région d'étiquette j ,
 - 3 : Identifier le couple d'étiquettes (i^*, j^*) tel que $P(i^*, j^*) = \max_{i,j} P(i, j)$,
 - 4 : Extraire toutes les régions d'étiquette i^* adjacentes à des régions d'étiquette j^* ,
 - 5 : Pour chacune des paires de régions adjacentes d'étiquette i^* et j^* , décider de les associer suivant un certain critère tenant compte du poids de l'arête.
 - 6 : Attribuer une même nouvelle étiquette aux régions issues des regroupements,
 - 7 : Retourner à l'étape 1, tant que les groupements effectués sont significatifs.
-

La matrice de probabilités $P_{i,j}$ peut juste compter les occurrences des paires de régions adjacentes d'étiquette i et j , ou introduire une pondération en tenant compte du degré d'adjacence entre les différents couples de régions. Par ailleurs, la décision d'associer une région de la classe i^* à une région de la classe j^* se fait suivant un certain critère dépendant du poids de l'arête. Par exemple, si le poids de l'arc qui les relie est supérieur à ceux des autres arêtes ayant l'une des deux régions pour extrémité, ou plus généralement, si le poids de l'arc entre les deux régions est supérieur à un certain seuil. En modifiant le critère associé au poids des arcs et la valeur du seuil, différents regroupements peuvent être envisagés. En particulier, si l'utilisateur peut interagir avec ces paramètres (soit directement, soit par apprentissage à travers des exemples), les structures mises en évidence n'en seront que plus pertinentes.

6.2.2 Détection des bâtiments et de leurs ombres

Dans cette section, nous nous focalisons sur le problème actuel de la détection des bâtiments et des ombres associées. Cette problématique est présente, entre autres, dans les tâches de reconstruction 3D des bâtiments, de mesure du degré d'urbanisation d'une commune ou encore de mise à jour de bases de données 2D. Dans la littérature, grand nombre des méthodes de détection des bâtiments sont basées sur la morphologie mathématique, mais peuvent cependant nécessiter une intervention forte de l'utilisateur dans le processus. Par exemple, dans [Matti-Gallice & Collet, 2004], les auteurs appliquent une classification supervisée en post-traitement des résultats fournis par les opérateurs morphologiques, pour l'extraction du bâti en milieu périurbain. D'autres méthodes réclament des données spécifiques telles que les modèles numériques d'élevation (MNE). Ainsi, disposant de modèles numériques d'élevation haute résolution, Ortner et ses collègues

[Ortner et al., 2003] modélisent un bâtiment par une silhouette rectangulaire, une fonction de coût indiquant la pertinence de la silhouette proposée, et un estimateur de toit permettant de construire un toit pour une silhouette pertinente donnée.

Nous proposons ici, une méthode non supervisée qui permet de détecter directement, dans une image satellitaire à haute résolution, des bâtiments et les ombres qui leur sont associées.

Nous avons effectué des tests sur une image Pelican à 50 cm de résolution, visualisée en fausses couleurs dans la figure 6.5 en haut à gauche. Cette image représentant le tissu urbain de la ville de Toulouse a tout d'abord été segmentée par l'algorithme Mean shift, puis les caractéristiques spectrales et géométriques de régions élémentaires obtenues ont subi une quantification vectorielle, suivant la procédure précédemment décrite. Le graphe d'adjacence des régions de l'image est pondéré par le rapport du nombre de pixels communs et du périmètre de la région formée en fusionnant les deux régions élémentaires :

$$\eta_{a,b} = \frac{\xi_{a,b}}{\pi_a + \pi_b - 2.\xi_{a,b}} \quad (6.6)$$

où π_a et π_b sont les périmètres des régions R_a et R_b respectivement, et $\xi_{a,b}$ est la longueur de la frontière commune entre ces deux régions.

Lors du premier balayage, les deux mots du vocabulaire visuel les plus souvent associés, c'est-à-dire maximisant la matrice des probabilités d'adjacence, correspondent à deux classes d'objets que nous sommes en mesure de nommer : il s'agit des bâtiments et des ombres, représentés en blanc et en marron (figure 6.5 en bas à gauche). Cependant, nous pouvons constater qu'il n'est pas possible de grouper systématiquement toutes les ombres adjacentes aux bâtiments. En effet, on y trouve les ombres des grands arbres, qui ne peuvent être associés aux bâtiments, mais aussi une ombre peut être adjacente à deux bâtiments proches. Le choix du seuil permet alors de prendre une décision par rapport aux deux régions à être groupées. Ici, le critère de regroupement tient compte de la compacité, c'est-à-dire de la régularité de la forme de la région obtenue, déterminée à l'aide du rapport isopérimétrique :

$$C_M = \frac{4\pi S}{P^2} \quad (6.7)$$

où S et P sont respectivement la surface et le périmètre de la région finale. La compacité est comprise entre 0 et 1, et est maximale pour un cercle. La compacité de la région obtenue par fusion de deux régions à grouper doit donc être supérieure à un seuil. En cas de conflit, si par exemple il est possible de grouper une région "ombre" avec chacune des deux régions "bâtiments" qui lui sont adacentes, on groupe l'ombre avec le bâtiment qui, fusionné à l'ombre, donne une région dont la forme est plus régulière. Dans la figure 6.5 en bas à gauche, les régions encerclées sont des exemples de bâtiments associés automatiquement à leurs ombres, même pour de fortes valeurs du seuil.

De manière itérative, il est possible de détecter aussi les grands arbres et leurs ombres, puis les bâtiments et les ombres avec les pelouses voisines, etc. C'est une hiérarchie de l'information pouvant mettre en évidence des structures sémantiques appartenant à des classes de mélange.

6.3 Apprentissage supervisé des relations entre les mots visuels

Dans cette section, nous revenons sur un problème rencontré au chapitre 4, pour la reconnaissance des classes de mélange de CORINE Land Cover, telles que les *zones industrielles*.

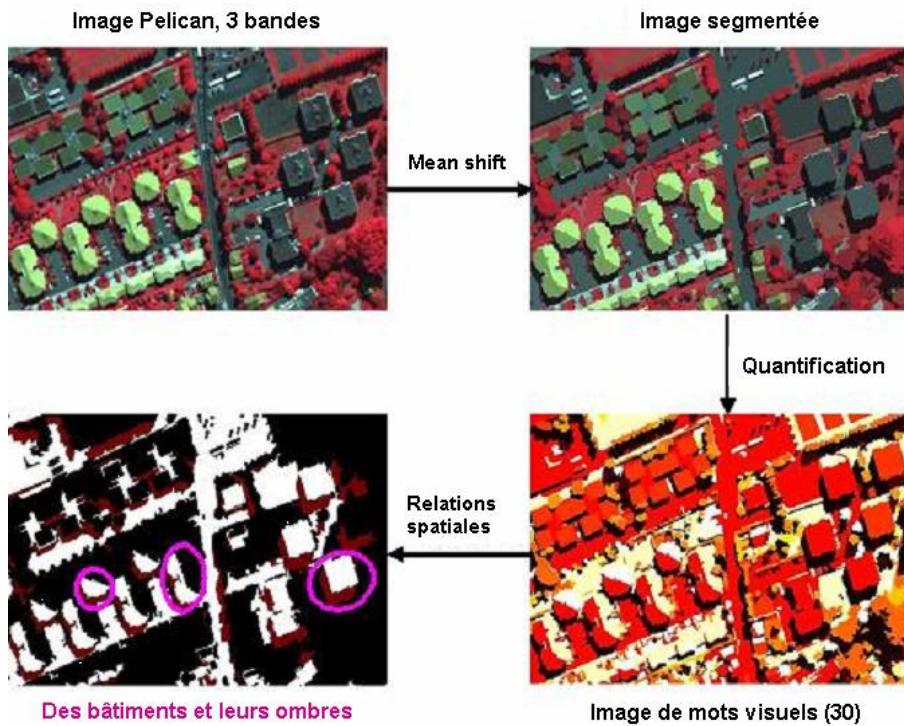


FIG. 6.5 – Exemple d'utilisation des relations d'adjacence entre les mots visuels pour la détection non supervisée des bâtiments et des ombres. Les "bâtiments" et les "ombres" ont été obtenus en exploitant la matrice de probabilité d'adjacence des mots visuels. Les régions encerclées sont des exemples de bâtiments associées à leurs ombres, même pour de fortes valeurs du seuil.

trielettes ou commerciales ou les forêts mélangées. En effet, un algorithme de classification basé sur le pixel n'était pas apte à identifier ce genre de classes, qui sont en fait des regroupements d'autres types de couverture et d'utilisation des terres. Nous proposons donc ici d'introduire une analyse au niveau de la région, et d'exploiter le principe de compositionnalité sémantique pour améliorer l'identification de telles classes. En effet, ces dernières peuvent être considérées comme des structures dans lesquelles chaque région représentée par un mot visuel est un composant de la classe de mélange.

6.3.1 Méthodologie

La méthodologie que nous avons suivie pour cet apprentissage est la suivante :

1. Après la procédure de création des mots visuels par segmentation de l'image suivie d'une quantification, nous identifions les mots visuels appartenant aux classes pures et construisons le graphe pondéré d'adjacence des régions. Il est également possible de partir d'une classification régularisée contextuellement, telle que proposée au chapitre 4. Dans ce cas, la taille du vocabulaire de mots visuels correspond au nombre de classes de l'image : chaque région a donc une étiquette.
2. Nous nous intéressons aux sous-graphes constitués :
 - des zones de taille inférieure à la taille minimale tolérée par CLC,

- des régions connexes à ces zones.
3. Nous considérons uniquement les noeuds dont l'étiquette est une étiquette d'une classe pure pouvant contribuer à une classe de mélange par exemple :
 - pour la classe 313 (*Forêts mélangées*), 311 (*Forêts de feuillus*) ou 312 (*Forêts de conifères*) ;
 - pour la classe 243 (*Systèmes culturaux et parcellaires complexes* : 211 (*Terres arables*), 211 (*Vignobles*), 231 (*Prairies*)) ;
 - pour la classe 112 (*Tissu urbain discontinu*) : 111 (*Tissu urbain continu*), 141 (*Tissu urbain verts*) et 142 (*Equipements sportifs et de loisirs*).
 4. Pour ces noeuds, nous examinons la façon dont ils ont été associés dans la zone d'apprentissage de CLC en prenant en compte deux choses :
 - (a) les étiquettes de l'image recouvertes par une étiquette unique de CLC,
 - (b) les étiquettes de l'image qui sont laissées à l'extérieur de la zone de mélange de CLC.
 5. Travaillant sur toutes les zones connexes d'intérêt du graphe, et cumulant ces associations et non-associations, nous essayons de tirer une règle de décision statistique.

6.3.2 Expérimentations sur les *forêts mélangées*

Sans aucune perte de généralité, nous nous intéressons dans la suite au cas particulier des *forêts mélangées* (classe 313), principalement composées des *forêts de feuillus* et des *forêts de conifères*. A partir de l'image de régions étiquetées (classification), il s'agit d'apprendre les règles de regroupement sur le graphe pondéré d'adjacence des régions, ceci en exploitant les règles de base de CORINE. En effet, l'apprentissage doit être effectué sous la contrainte du respect de la taille de la région minimale dans CORINE Land Cover, et de sa nomenclature, qui définit les *forêts mélangées* comme des *formations végétales principalement constituées par des arbres, mais aussi par des buissons et des arbustes, où ni les feuillus, ni les conifères ne dominent*. Ensuite, forts de cet apprentissage, nous allons essayer de retrouver les zones de *forêts mélangées* dans les images.

Les regroupements des régions sur le graphe sont effectués par croissance de régions. Partant d'une région initiale appartenant soit aux *forêts de feuillus*, soit aux *forêts de conifères*, le graphe est parcouru en *largeur d'abord*. L'algorithme de parcours en largeur d'abord découvre tous les sommets à distance k avant de découvrir un sommet à distance $k + 1$. Ayant découvert tous les sommets à distance $k - 1$ et une partie des sommets à distance k , on commence par découvrir tous les voisins des sommets à distance $k - 1$ avant de commencer à considérer les voisins des sommets à distance k .

A chaque étape du parcours du graphe, le coût de regroupement des deux régions est calculé. La fonction de coût utilisée est conçue en tenant compte au minimum des contraintes de superficie et de proportion de chacune des composantes dans une zone de *forêts mélangées*. En effet, on vérifie les conditions suivantes :

- La superficie de la région obtenue par regroupement est supérieure à la superficie de la région minimale dans CORINE Land Cover, soit $s_{min} = 25ha$.
- Les proportions de forêts de feuillus et de conifères dans la nouvelle région sont respectées. Soient α_{min} et α_{max} , les bornes inférieure et supérieure de la proportion de feuillus ou de conifères devant être présentes dans la région, soient β_f et β_c , les

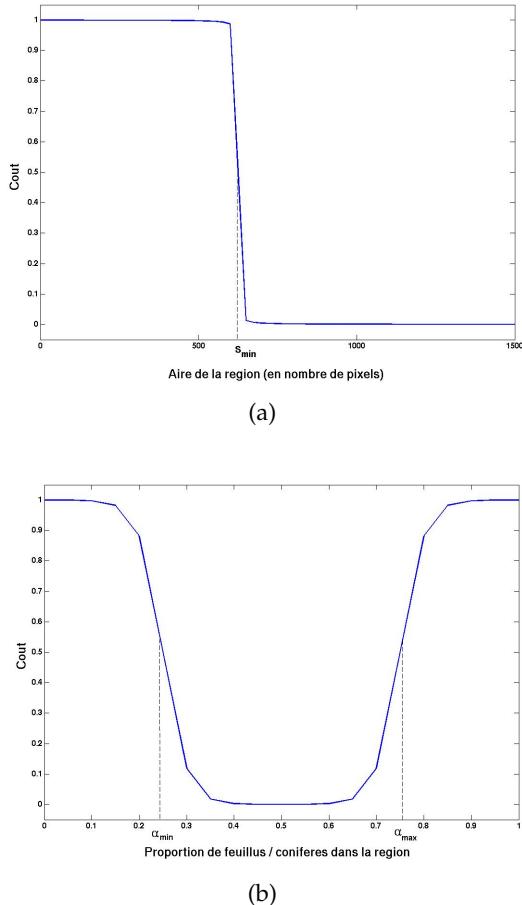


FIG. 6.6 – Fonctions de coût de regroupement associées à la contrainte de superficie (a) et à la contrainte de proportion des feuillus ou de conifères (b) dans la région obtenue.

proportions effectives de feuillus et de conifères présentes dans la nouvelle région, alors :

$$\begin{cases} \alpha_{min} \leq \beta_f \leq \alpha_{max} \\ \alpha_{min} \leq \beta_c \leq \alpha_{max} \\ \text{et } \beta_f + \beta_c \leq 1 \end{cases}$$

Les différents coûts C_p et C_s associés à chaque contrainte peuvent s'exprimer à l'aide des équations suivantes par exemple :

$$\begin{aligned} C_p &= \max \left(1 - f(\beta_f - \alpha_{min}) + f(\beta_f - \alpha_{max}), 1 - f(\beta_c - \alpha_{min}) + f(\beta_c - \alpha_{max}) \right) \\ C_s &= \frac{1}{\pi} \cdot \left(\arctan(-(s - s_{min})) + \frac{\pi}{2} \right) \end{aligned} \quad (6.8)$$

où f est une fonction sigmoïde telle que :

$$f(x) = \frac{1}{1 + \exp(\lambda x)}$$

Ces contraintes sont représentées dans les figures 6.6(a) et 6.6(b). Le coût total s'exprime alors par : $C = a.C_s + b.C_p$, les coefficients a et b exprimant le poids accordé à



FIG. 6.7 – Image SPOT de la région du Loiret, 20m de résolution et 3 bandes spectrales : R, V et PIR. Les zones (1 à 5) détournées en blanc sont des *forêts mélangées* suivant la classification de CORINE Land Cover.

chaque contrainte.

Bien sûr, d'autres contraintes peuvent être prises en compte, à l'instar du nombre de noeuds moyen (régions élémentaires) dans la région obtenue, de la taille moyenne des noeuds, etc.

Sur les zones de *forêts mélangées* utilisées pour l'apprentissage, le graphe est donc parcouru, les relations spatiales entre les régions sont apprises de cette manière, en calculant la fonction de coût.

Sur une image test décomposée en mots visuels, le graphe est parcouru à partir d'une région initiale. Et pour chaque regroupement, la fonction de coût est calculée. Si le coût obtenu est supérieur à un seuil déterminé lors de l'apprentissage, l'algorithme arrête les regroupements pour cette région d'initialisation et passe à une autre région pour initialiser un nouveau regroupement ou une nouvelle structure.

La figure 6.7 montre une image SPOT de la région du Loiret dans laquelle les zones de *forêts mélangées* selon CORINE Land Cover sont détournées (zones 1 à 5). Les tests effectués avec la méthode décrite ci-dessus ont donné les résultats de la figure 6.8. Les zones détectées comme étant des *forêts mélangées* (détournées en vert) peuvent être classées en deux catégories : celles qui ont une intersection avec une *forêt mélangée* de CORINE, et les fausses alarmes.

Dans le premier cas, les délimitations des zones détectées comme étant des *forêts mélangées* ne sont pas les mêmes que celles données par CORINE. Par exemple, lors des tests de regroupement, certaines régions de la zone 2 ont été associées à d'autres régions pour former une *forêt mélangée* recouvrant partiellement la zone 2. Pour évaluer l'identification des zones de *forêt mélangée* données par la vérité de terrain, nous avons utilisé la



FIG. 6.8 – Résultats donnés par les tests de regroupement : les zones détournées en vert représentent les *forêts mélangées* obtenues.

TAB. 6.1 – Evaluation de la reconnaissance par regroupement des régions élémentaires, des zones de *forêts mélangées* délimitées par la classification de CORINE Land Cover.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
$\frac{N_{com}}{N_v}$	0.50	0.47	0	0.80	0.34
$\frac{N_{com}}{N_c}$	0.84	0.48	0	0.51	0.36
t_R	0.42	0.23	0	0.40	0.12
$F - score$	0.63	0.48	0	0.61	0.34

quantité t_R , définie comme suit :

$$t_R = \frac{N_{com}}{N_v} \cdot \frac{N_{com}}{N_c} \quad (6.9)$$

où N_{com} est la taille (nombre de pixels) de la zone commune entre la *forêt mélangée* réelle et celle détectée, N_v , la taille de la *forêt mélangée* selon la vérité de terrain, et N_c , la taille de la *forêt mélangée* obtenue. t_R est égale à 1 lorsque la zone détectée après regroupement des régions est identique à celle de la vérité de terrain, et nulle lorsqu'il n'y a aucune intersection. La grandeur $\frac{N_{com}}{N_v}$ est appelée *rappel*, tandis que $\frac{N_{com}}{N_c}$ est la *précision*. Une mesure populaire qui combine la précision et le rappel est leur pondération, nommée F-score :

$$F - score = 2 \cdot \frac{\text{precision} \times \text{rappel}}{\text{precision} + \text{rappel}} \quad (6.10)$$

La table 6.1 montre les résultats obtenus pour les 5 régions. On note que la région 1 est la mieux identifiée.



FIG. 6.9 – Exemple de zone détectée comme forêt mélangée (détourée en vert), par regroupement de régions élémentaires. La zone détourée en blanc est la forêt mélangée correspondant à la vérité de terrain CLC. La région initiale a été choisie manuellement.

Les frontières des régions obtenues varient en fonction de la région initiale choisie pour le parcours du graphe. Lorsque la région est choisie de manière judicieuse, il est possible d'obtenir une *forêt mélangée* plus proche de celle de CORINE. La figure 6.9 montre un exemple de *forêt mélangée* obtenue (en vert) tandis que dans la vérité de terrain, la *forêt mélangée* est la zone détourée en blanc. Dans ce cas, t_R vaut 0.46 (au lieu de 0.23). Notons que les contours des deux zones ne coïncident toujours pas. Ceci peut être dû au fait que la classification de CORINE est manuelle et peut être remise en cause.

En ce qui concerne les fausses alarmes, ces zones détectées comme étant des forêts mélangées satisfont les contraintes de CORINE. En effet, chacune de ces zones respecte le critère de superficie minimale et répond à la définition d'une *forêt mélangée* donnée par la nomenclature. Dans la figure 6.8, les régions rouges correspondent aux *forêts de feuillus*, tandis que les régions sombres sont des *forêts de conifères*. La plupart des *forêts mélangées* détectées contiennent en majorité ces deux classes. Ceci nous amène à penser qu'il existe des informations utilisées par la classification de CORINE dont nous n'avons pas tenu compte. Ces informations pourraient être présentes dans l'image, mais surtout extérieures à l'image satellitaire, d'où l'importance des données exogènes dans la constitution des cartes CORINE Land Cover.

6.4 Conclusions

Dans ce chapitre, nous illustrons le principe de compositionnalité sémantique que nous avons appliqué aux images. En effet, l'importance de l'utilisation des relations spatiales entre les mots visuels pour identifier des classes de mélange est mise en évidence. Nous avons tout d'abord insisté sur le choix du mode de représentation de ces relations spatiales, important pour exploiter efficacement les liens d'adjacence entre les mots visuels. Deux approches ont été présentées, faisant usage des graphes pour manipuler les relations d'adjacence : une approche non supervisée nous permettant de déterminer les

structures dans l'image, exemplifiée par la détection des bâtiments et leurs ombres dans les images satellitaires, et une approche supervisée d'apprentissage des relations spatiales, testée sur la classe des forêts mélangées de CORINE Land Cover. Dans le chapitre suivant, nous nous attachons à résoudre notre problème à l'aide de modèles statistiques d'analyse textuelle, qui ne prennent pas en compte les relations spatiales entre les mots visuels.

Chapitre 7

Annotation sémantique des images satellitaires basée sur le modèle LDA

Dans cette partie, nous mettons en oeuvre un apprentissage sémantique supervisé basé sur un modèle faisant abstraction des relations d'adjacence entre les mots visuels, afin d'attribuer aux régions de l'image, une sémantique dépendant du contexte de l'utilisateur. Nous l'illustrons via une application très présente dans la littérature, et qui est encore l'objet de plusieurs travaux (voir section 5.4.1) : l'annotation sémantique d'images. Différentes approches ont été proposées pour annoter les images naturelles. Ici, nous nous intéressons aux images satellitaires et proposons d'utiliser une méthode d'analyse de textes : l'Allocation Dirichlet Latente (Latent Dirichlet Allocation ou LDA) [Blei et al., 2003b] pour l'annotation de grandes images satellitaires, en utilisant des concepts sémantiques définis par l'utilisateur. L'approche présentée ici combine une étape de classification supervisée d'imagettes extraites de la grande image à annoter, et une phase de prise en compte de l'information spatiale entre ces imagettes.

7.1 Approche d'annotation sémantique de grandes images

Dans cette partie, nous décrivons notre approche d'annotation sémantique de grandes images satellitaires, exploitant le modèle LDA.

Soient S concepts sémantiques définis par l'utilisateur. Pour chaque concept, ce dernier fournit un ensemble d'images exemples qui seront utilisées pour l'apprentissage. Soit une grande image I à annoter avec les S concepts sémantiques, nous appellerons ensemble de test, la collection des imagettes I_d , de taille égale, telles que :

$$\bigcup_d I_d = I \quad (7.1)$$

L'annotation de la grande image I sera considérée pour nous, comme la classification supervisée des imagettes I_d en S classes $C_s, s \in \{1, \dots, S\}$ correspondant aux concepts sémantiques définis par l'utilisateur. La classification supervisée est basée sur le modèle LDA proposé dans [Blei et al., 2003b]. Par définition, l'Allocation Dirichlet Latente est un modèle génératif probabiliste pour les collections de données discrètes telles que les corpus de texte. Comme détaillé dans la section 5.4.2.2, l'idée de base de ce modèle hiérarchique à trois niveaux, basé sur une représentation en sac-de-mots, est que les documents d'un corpus sont représentés comme des mélanges sur des topics latents, chaque

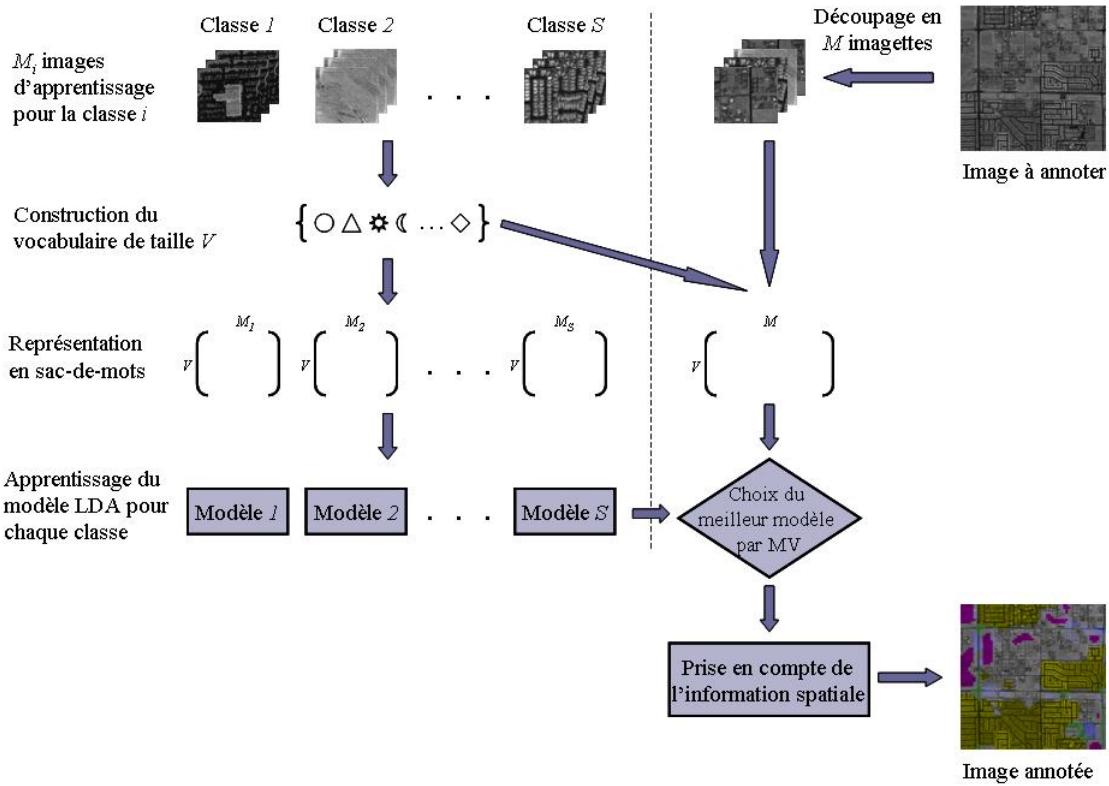


FIG. 7.1 – Les différentes étapes de notre algorithme d'annotation sémantique d'images de grande taille.

topic étant caractérisé par une distribution sur des mots. Dans la section 5.4.2.3, nous avons détaillé le processus d'utilisation du LDA pour la modélisation des images. Notre tâche de classification inclut donc le calcul des mots visuels pour la représentation des images, la génération du modèle pour chaque classe lors de l'apprentissage, et l'application sur les images de l'ensemble de test.

Par ailleurs, l'ordre des images dans une collection étant négligé dans le modèle LDA, nous introduisons l'information spatiale entre les imagettes de l'ensemble de test, de manière à prendre en compte les relations entre des imagettes voisines dans la grande image. Nous montrerons par la suite, à l'aide d'expérimentations effectuées sur une série d'images, que cette opération améliore la tâche d'annotation de manière notable.

Les différentes étapes de notre approche d'annotation sont récapitulées dans la figure 7.1.

7.1.1 Classification supervisée des imagettes

L'utilisation du modèle LDA pour classifier les images de la base de test en S classes sémantiques nécessite une étape de pré-traitement des données. En effet, disposant d'un ensemble d'images d'apprentissage fournies par l'utilisateur pour chaque classe, et d'un ensemble d'images de test, la première étape consiste à calculer les mots visuels pour la représentation en sac-de-mots des images de la base de données. Suivant la procédure décrite dans la section 5.4.2.3 pour la construction d'un vocabulaire visuel, chaque image

de la base (apprentissage et test) est décomposée en mots, d'où sont automatiquement extraites des primitives. La quantification vectorielle sur l'ensemble des primitives de la base d'apprentissage pour la construction du vocabulaire, est effectuée via l'algorithme classique d'agrégation de données (clustering) k-means. Afin de déterminer le nombre optimal de clusters, nous utilisons une approche décrite dans [Kyrgyzov et al., 2007], qui modélise les descripteurs par un mélange de gaussiennes, et utilise le critère de la longueur de description minimale (*Minimal Description Length* ou MDL) pour accéder à la complexité optimale du modèle. La taille du vocabulaire V correspond au nombre optimal de clusters, dont les centres représentent les mots visuels.

Une fois les images représentées comme des séquences de mots visuels, le modèle de génération des images d'apprentissage pour chaque classe est obtenu par le LDA. Chaque image étant modélisée comme un mélange de topics latents appris de manière non supervisée, il est au préalable nécessaire de déterminer le nombre de topics pour chaque classe. Une mesure appropriée pour évaluer la performance du modèle en fonction de ce paramètre est donc nécessaire. Utilisée par convention dans le domaine de la modélisation du langage, la *perplexité* est une mesure qui quantifie la capacité d'un modèle à prédire de nouvelles données. Pour un ensemble de test $D_{test} = \{\mathbf{w}_1, \dots, \mathbf{w}_M\}$ de M images, la perplexité est donnée par l'expression :

$$\text{perplexite}(D_{test}) = \exp \left(-\frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right) \quad (7.2)$$

Cette mesure est décroissante en fonction de la vraisemblance de l'ensemble de test, qui dépend du nombre de topics dans la collection. Une valeur faible de la perplexité est donc indicatrice d'une bonne performance de généralisation aux nouvelles données. Ainsi, pour chaque classe, différents modèles LDA sont appris en utilisant différents nombres de topics latents. La perplexité est ensuite calculée pour chaque modèle sur un ensemble d'images test, permettant ainsi d'exprimer la valeur de la perplexité en fonction du nombre d'aspects latents. Le nombre optimal de topics pour la classe est celui qui minimise la perplexité.

L'apprentissage basé sur le LDA utilise donc un nombre fini de topics latents pour générer un modèle représentant le mieux la distribution des mots visuels pour chaque classe.

Soit une image \mathbf{w} de l'ensemble de test, représentée par une séquence de mots visuels, nous souhaitons calculer la probabilité de chaque classe C_s , $s \in \{1, \dots, S\}$: $p(C_s | \mathbf{w}, \alpha_s, \beta_s)$, où α_s et β_s sont les paramètres du modèle appris à partir de l'ensemble d'apprentissage de la classe sémantique C_s . Selon la formule de Bayes :

$$p(C_s | \mathbf{w}, \alpha_s, \beta_s) \propto p(\mathbf{w} | \alpha_s, \beta_s) \times p(C_s) \quad (7.3)$$

En général, la distribution de $p(C_s)$ est supposée uniforme et fixée, et s'exprime dans notre cas par $p(C_s) = 1/S$. Il en résulte que :

$$p(C_s | \mathbf{w}, \alpha_s, \beta_s) \propto p(\mathbf{w} | \alpha_s, \beta_s) \quad (7.4)$$

où $p(\mathbf{w} | \alpha_s, \beta_s)$ est la vraisemblance de l'image pour la classe C_s , définie dans l'équation 5.3. En tenant compte de toutes les classes, on obtient ainsi un vecteur de probabilités

$[p(\mathbf{w}|\alpha_1, \beta_1) p(\mathbf{w}|\alpha_2, \beta_2) \dots p(\mathbf{w}|\alpha_S, \beta_S)]$ pour chaque image \mathbf{w} de l'ensemble de tests. La prise de décision repose sur la méthode du Maximum de Vraisemblance (MV) qui attribue l'image \mathbf{w} à la classe qui maximise la vraisemblance :

$$s^* = \operatorname{argmax}_s p(\mathbf{w}|\alpha_s, \beta_s) \quad (7.5)$$

7.1.2 Prise en compte de l'information spatiale

L'approche que nous avons adoptée pour l'annotation de grandes images est de classifier à l'aide du modèle LDA, des imagettes extraites de la grande et obtenues par découpage suivant une grille régulière. L'ordre des documents dans un corpus étant supposé non pertinent dans le modèle LDA, l'information spatiale entre des imagettes adjacentes est perdue lors de la classification. Afin de conserver cette information et l'exploiter pour améliorer l'annotation, nous proposons de découper la grande image en imagettes de taille $d \times d$ suivant un maillage régulier de pas p , avec $p < d$ de sorte qu'il y ait un recouvrement entre les imagettes. En d'autres termes, plusieurs imagettes de l'ensemble de test peuvent avoir des zones en commun, que nous appellerons *patches*, de taille $p \times p$. Cela introduit une redondance qui est ensuite exploitée après le processus de classification basé sur le LDA. En effet, des imagettes voisines peuvent être attribuées à des classes différentes, aboutissant ainsi à plusieurs possibilités de classes pour les patches qu'elles partagent. Pour remédier à cela, un vote majoritaire est opéré sur chaque patch, en l'attribuant à la classe la plus représentée dans l'ensemble des différentes classes possibles pour ce patch. Au final, cela revient à classifier les patches en S classes sémantiques, en exploitant les relations de voisinage. Les patches étant de taille plus petite que les imagettes, l'annotation de la grande image est plus fine.

7.2 Expérimentations et résultats

Plusieurs séries d'expérimentations ont été menées pour tester l'approche sus-décrise. En fait, deux bases d'images satellitaires, à des résolutions différentes, nous ont permis de juger de l'intérêt de cette méthode :

- les images Quickbird Panchromatiques de Las Vegas (USA), à 60 cm de résolution : avec cette base, nous évaluons l'influence de l'information spatiale dans notre approche d'annotation d'images, en effectuant deux types de tests : sans et avec prise en compte des relations de voisinage entre les imagettes.
- les images Quickbird Multispectrales de Marseille (France), à 2.44 m de résolution : ici, les caractéristiques radiométriques ainsi que des primitives de texture sont prises en compte.

Pour chaque base d'images, les images exemples pour l'apprentissage, fournies par l'utilisateur pour chaque classe n'appartiennent pas à l'image test à annoter.

En outre, dans toutes les expérimentations présentées dans cette section, les mots et les documents sont des fenêtres de pixels, dont les tailles dépendent de la résolution des images et du niveau de sémantique des concepts définis pour chaque base, comme nous le verrons lors des tests.

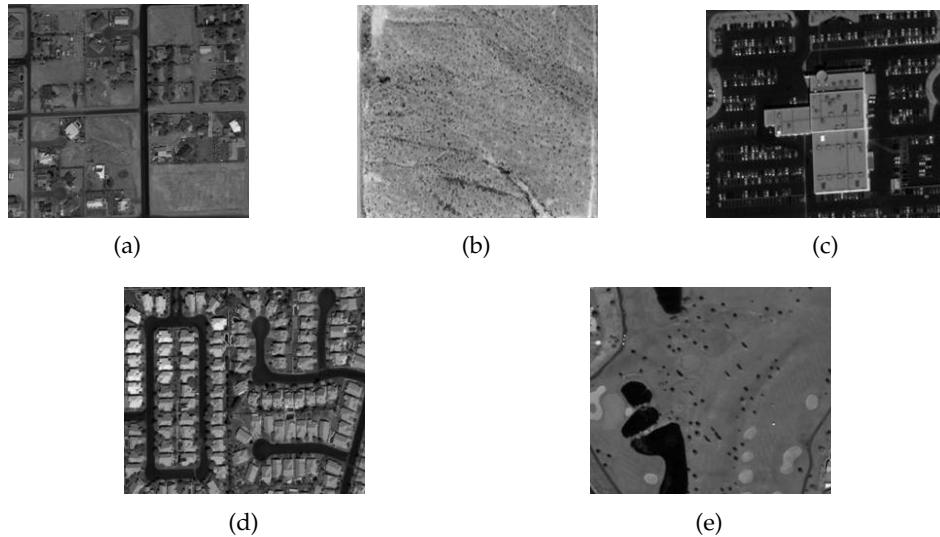


FIG. 7.2 – Exemples d’images de la base d’apprentissage Quickbird de Las Vegas, appartenant aux cinq classes reliées aux concepts sémantiques utilisés dans nos expérimentations. (a) : banlieues résidentielles (BR), (b) : déserts (DS), (c) : zones commerciales (ZC), (d) : zones urbaines (ZU), (e) : terrains de golf (TG).

7.2.1 Influence de l’information spatiale

7.2.1.1 Description des données

Les tests sont effectués sur des images Quickbird Panchromatiques de Las Vegas, à 60 cm de résolution. La base d’apprentissage fournie par l’utilisateur contient 200 images de taille variant de 150×150 à 500×500 pixels. Ces images, dont quelques exemples sont exposés dans la figure 7.2, sont également réparties entre 5 classes correspondant aux concepts suivants : banlieues résidentielles (BR), déserts (DS), zones commerciales (ZC), zones urbaines (ZU) et terrains de golf (TG).

La grande image à annoter avec les concepts que nous venons de définir est de taille 6000×6000 pixels (voir figure 7.3). Elle contient toutes les classes sus-présentées, mais aussi d’autres classes non apprises, comme les axes routiers principaux.

7.2.1.2 Réglage des paramètres et tests

Dans cette expérimentation, un mot correspond à une fenêtre de taille 10×10 pixels, déterminée de manière empirique. En effet, la taille du mot doit être suffisamment grande pour contenir au minimum une partie d’un objet de l’image, et suffisamment petite pour contenir une information homogène, c’est-à-dire ne pas mêler des informations appartenant à des objets différents par exemple. Diverses expérimentations ont donc été menées avec des fenêtres de taille 10×10 , 15×15 et 20×20 pixels, et la décision s’est faite sur les performances des tests de validation croisée. Nous avons utilisé une procédure de validation croisée à 5 boucles : l’ensemble des données de la base d’apprentissage est divisé de manière aléatoire en 5 sous-ensembles. Ainsi, 20% des images sont utilisées pour tester l’apprentissage fait à partir de 80% de la population, chaque sous-ensemble étant



FIG. 7.3 – Image Quickbird de Las Vegas à annoter.

utilisé une fois et une seule pour les tests. Les meilleures performances ont été obtenues avec les mots de taille 10×10 pixels.

Disposant d'images panchromatiques, les caractéristiques de bas-niveau automatiquement extraites de chaque mot de la base de données sont simplement la moyenne et l'écart-type. En fait, nous avons effectué des tests avec d'autres descripteurs tels que la médiane et la variation totale, et le couple de primitives retenu s'est avéré être le plus pertinent en termes de performance et de nombre de caractéristiques. Par exemple, l'utilisation simultanée de la moyenne, la variance et la variation totale n'améliore pas pour autant les performances des tests de validation croisée obtenues avec la moyenne et la variance.

Le nombre optimal de clusters pour nos données, déterminé à l'aide du critère MDL est de 20, comme le montre la figure 7.4. Quant au nombre de topics pour chaque classe, il est estimé grâce à la perplexité. La figure 7.5 montre la courbe de la perplexité en fonction du nombre de topics pour la classe des banlieues résidentielles. Le nombre optimal d'aspects latents est celui qui minimise la complexité, en l'occurrence, 5 dans le cas des banlieues résidentielles.

Les modèles sont évalués par une procédure de validation croisée à 5 boucles. Les performances des modèles des différentes classes sont détaillées dans la matrice de confusion exposée dans le tableau 7.1. Nous constatons qu'il existe quelques confusions justifiables entre certaines classes. En effet, des images de la classe *déserts* par exemple sont classifiées en tant que *terrains de golf*, et vice versa. Ce qui n'est pas vraiment étonnant

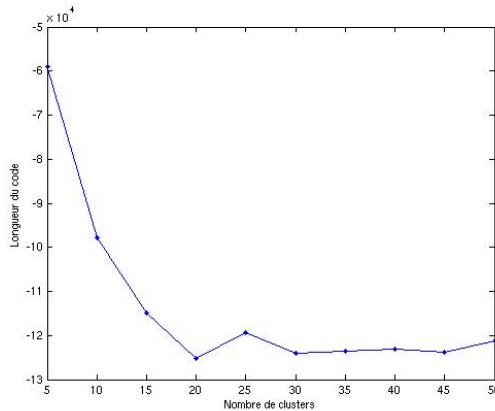


FIG. 7.4 – Courbe de la longueur du code en fonction du nombre de clusters pour l’ensemble des données. Le nombre optimal de clusters est celui qui minimise longueur de description, c’est-à-dire 20 dans ce cas.

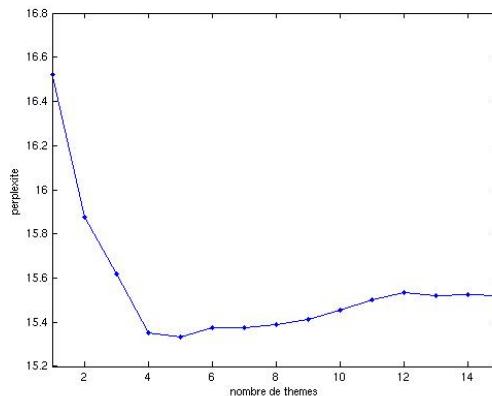


FIG. 7.5 – Courbe de la perplexité en fonction du nombre de topics pour la classe des banlieues résidentielles. Le nombre de topics optimal est celui qui minimise la perplexité, c’est-à-dire 5 dans ce cas.

puisque les terrains de golf sont principalement composés de champs, déserts et lacs.

L’ensemble de test contient des imagettes issues de la grande image (figure 7.3), et de taille 150×150 pixels choisie expérimentalement. Chaque imagette contient donc 225 mots. Nous avons effectué deux types de tests sur l’image à annoter : sans et avec recouvrement entre les imagettes. Nous comparons ensuite les deux images annotées résultantes, afin de juger de l’apport de la prise en compte de l’information spatiale. L’évaluation des résultats est visuelle, en utilisant les cartes Google comme vérité de terrain.

7.2.1.3 Analyse des résultats

La figure 7.6 représente l’image de test, annotée avec les cinq concepts sémantiques et sans prise en compte de l’information de voisinage. En d’autres termes, les imagettes de l’ensemble de test sont obtenues par partitionnement de l’image à annoter, selon une

TAB. 7.1 – Matrice de confusion obtenue pour la validation croisée à 5 boucles. Les lignes représentent la vérité de terrain et les colonnes correspondent aux modèles des différentes classes. La performance moyenne est de 96.5%.

	BR	DS	ZC	ZU	TG
BR	97.5	0	0	2.5	0
DS	2.5	95.0	0	0	2.5
ZC	0	0	97.5	2.5	0
ZU	2.5	0	0	97.5	0
TG	2.5	5.0	0	0	92.5

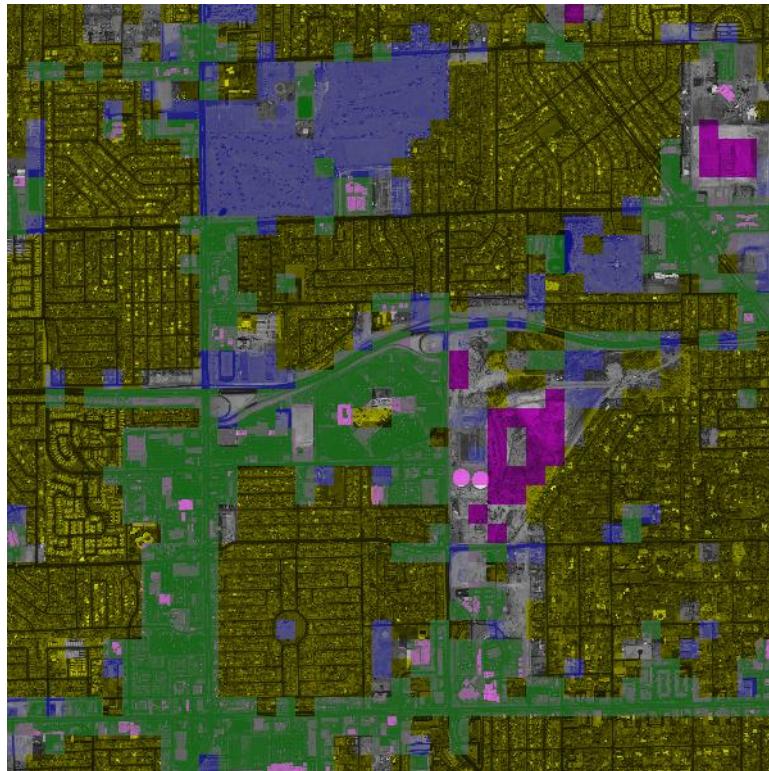


FIG. 7.6 – Image annotée sans prise en compte de l’information de voisinage, en utilisant les 5 concepts sémantiques suivants : déserts (rose), zones commerciales (vert), zones urbaines (jaune), terrains de golf (bleu) et banlieues résidentielles (sans couleur).

grille régulière dont le pas est égal à la taille des imagettes. L’ensemble de test contient donc 40×40 imagettes, à classifier en cinq classes. De manière globale les résultats sont intéressants, d’autant plus que les descripteurs utilisés pour le calcul des mots visuels sont sobres. Nous constatons cependant des confusions entre les classes *déserts* et *banlieues résidentielles*. Ceci est dû au fait que les *banlieues résidentielles* sont une classe de mélange, principalement composée de maisons, déserts et de petits espaces verts. En outre, plu-

sieurs régions mal annotées sont celles qui ne correspondent à aucun des concepts définis par l'utilisateur. Par exemple, les routes principales sont attribuées aux *zones commerciales*, tandis que les pelouses sont classifiées comme étant des *terrains de golf*. Rappelons que les caractéristiques bas-niveau utilisées pour ces expérimentations sont simplement la moyenne et l'écart-type, et la décision pour la classification est faite par Maximum de Vraisemblance. Donc ces erreurs d'annotation ne sont guère surprenantes. Pour y remédier, nous projetons d'introduire une classe de rejet pour les zones ne correspondant à aucun des concepts utilisés, ou alors de définir des concepts sémantiques couvrant toutes les régions possibles de l'image de test. De même, des descripteurs locaux plus riches pourraient largement réduire les erreurs d'annotation.

Dans la seconde expérimentation, les imagettes de l'ensemble de test sont obtenues par découpage de l'image à annoter avec un recouvrement de 50 pixels. Ainsi, après la classification des 118×118 imagettes de l'ensemble de test en 5 classes, un vote majoritaire est mis en oeuvre pour reconstruire l'image, en tenant compte des relations de voisinage entre les imagettes. Les résultats semblent satisfaisants, comme le montre la figure 7.7. Par rapport au test précédent, l'amélioration de l'annotation est flagrante : il y a moins d'imagettes isolées puisque les relations de voisinage sont introduites, et les zones de l'image appartenant aux différents concepts sont moins grossières et mieux délimitées. En effet, le grand terrain de golf en haut de l'image par exemple, correspond mieux à la vérité de terrain avec ce second test.

Par conséquent, dans la suite, nous ne présenterons que les résultats des expérimentations qui tiennent compte de l'information de voisinage.

7.2.2 Utilisation des caractéristiques radiométriques et texturelles

7.2.2.1 Description des données

Les images Quickbird de Marseille à notre disposition sont multispectrales à 4 canaux (rouge, vert, bleu et infrarouge). La base d'apprentissage est constituée de 320 images de même taille, dont quelques exemples sont présentés dans la figure 7.9. Les concepts suivants sont utilisés pour les expérimentations : carrières (CA), espaces verts (EV), grands bâtiments (GB), mer (ME), montagne (MT), ports (PO), zones résidentielles (ZR) et zones urbaines (ZU).

L'image de test, de taille 2400×3040 pixels, est assez complexe (figure 7.8). En effet, plusieurs petites régions, appartenant à des classes sémantiques différentes, sont mêlées les unes avec les autres, rendant ainsi épingleuse toute tentative d'annotation manuelle complète de l'image.

7.2.2.2 Protocole de tests

Les images de la base ayant une résolution de 2.44 m par pixel, un mot correspond ici à une fenêtre de taille 4×4 pixels. Les caractéristiques bas-niveau extraites des mots sont la moyenne pour chaque bande spectrale, l'indice de végétation NDVI, l'indice de brillance IB et l'indice de bâti ISU. Nous introduirons ensuite des primitives de texture, basées sur les filtres miroirs en quadrature (Quadrature Mirrors Filters ou QMF). Ces primitives sont extraites de fenêtres de taille 64×64 obtenues par découpage de l'image avec un recouvrement, puis interpolées de manière à attribuer un vecteur de caractéristiques de texture à chaque mot. Le nombre optimal de clusters pour le calcul des mots visuels

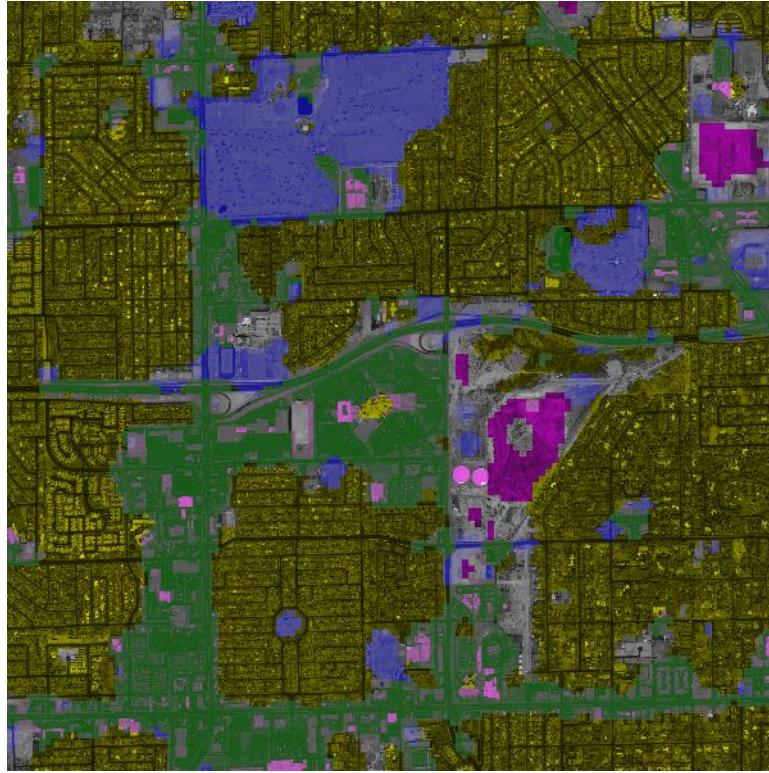


FIG. 7.7 – Image annotée avec prise en compte des relations de voisinage entre les imagettes, en utilisant les 5 concepts sémantiques suivants : déserts (rose), zones commerciales (vert), zones urbaines (jaune), terrains de golf (bleu) et banlieues résidentielles (sans couleur).

ainsi que le nombre de topics pour la génération du modèle LDA pour chaque classe sont déterminés de la même manière que dans l’expérimentation précédente.

Les performances des modèles sont évaluées par validation croisée, utilisant à chaque fois 80% de la base d’images pour l’apprentissage, et 20% pour les tests.

L’ensemble de test est composé d’imagettes de taille 32×32 pixels, issues du découpage de la grande image selon une grille régulière, avec un recouvrement de 16 pixels. Chaque imagette possède ainsi 64 mots. Nous nous proposons de réaliser une annotation de l’image de test en n’utilisant que des primitives spectrales dans un premier temps. Nous nous interrogerons ensuite sur la pertinence des caractéristiques de texture pour l’annotation sémantique d’une telle image.

7.2.2.3 Evaluation

Afin de juger de la qualité de l’annotation obtenue, une annotation manuelle de l’image de test peut être utilisée comme vérité de terrain. Cependant, l’annotation manuelle complète d’une grande image est une tâche très coûteuse, plus coûteuse encore pour notre image de test compte tenu de sa complexité. Aussi, nous nous proposons de déterminer manuellement et de manière aléatoire, un certain nombre de régions pour chaque classe, et de mesurer le degré de reconnaissance de chacune de ces régions par notre approche d’annotation. En d’autres termes, pour chaque région R_{si} appartenant à

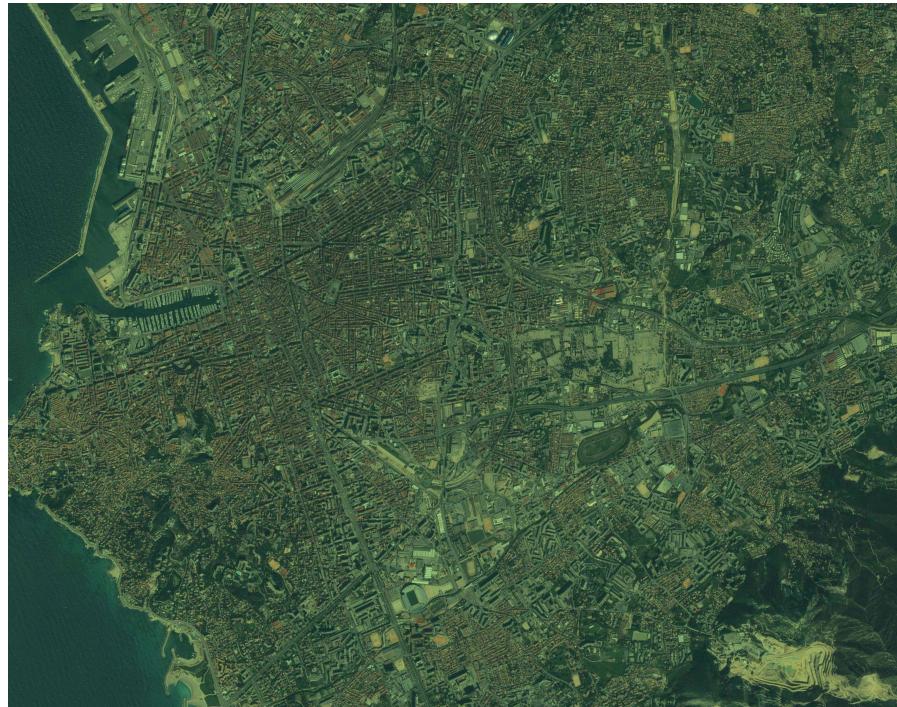


FIG. 7.8 – Image Quickbird de Marseille à annoter.

un concept sémantique s selon la vérité de terrain (annotation manuelle), soient N_{sm} le nombre de patches appartenant à cette région, et N_{sa} le nombre de patches de l'image annotée résultante, inclus dans la région R_{si} et associés au concept s , la région R_{si} est parfaitement identifiée lorsque le degré de reconnaissance d_R est proche de 1 :

$$d_R = \frac{N_{sa}}{N_{sm}} \quad (7.6)$$

Précisons que le nombre de régions annotées manuellement n'est pas le même pour toutes les classes étant donnée l'inégale répartition des différentes classes dans l'image de test. En outre, lorsque plusieurs régions de test ont été retenues pour un concept sémantique, ces dernières sont si possible choisies de telle sorte que leurs contextes spatiaux soient différents. Ceci dans le but de diversifier les tests afin d'avoir une évaluation plus complète, reflétant au mieux la réalité. Nous jugeons ainsi de la capacité de notre système d'annotation à reconnaître les régions de test indépendamment de leur voisinage. Enfin, pour chaque concept sémantique, les tailles des régions sont du même ordre de grandeur.

7.2.2.4 Analyse des résultats

Caractéristiques radiométriques Nous présentons tout d'abord les résultats obtenus en n'utilisant que les caractéristiques spectrales pour le calcul des mots visuels, c'est-à-dire les 4 bandes spectrales et les néocanaux. La validation croisée a donné en moyenne 95.31% de bonne classification, quelques confusions justifiées par la nature des descripteurs utilisés étant présentes. La structure hiérarchique de la figure 7.10 établit les relations entre les différentes classes. Ces liens sont évalués en calculant une distance entre

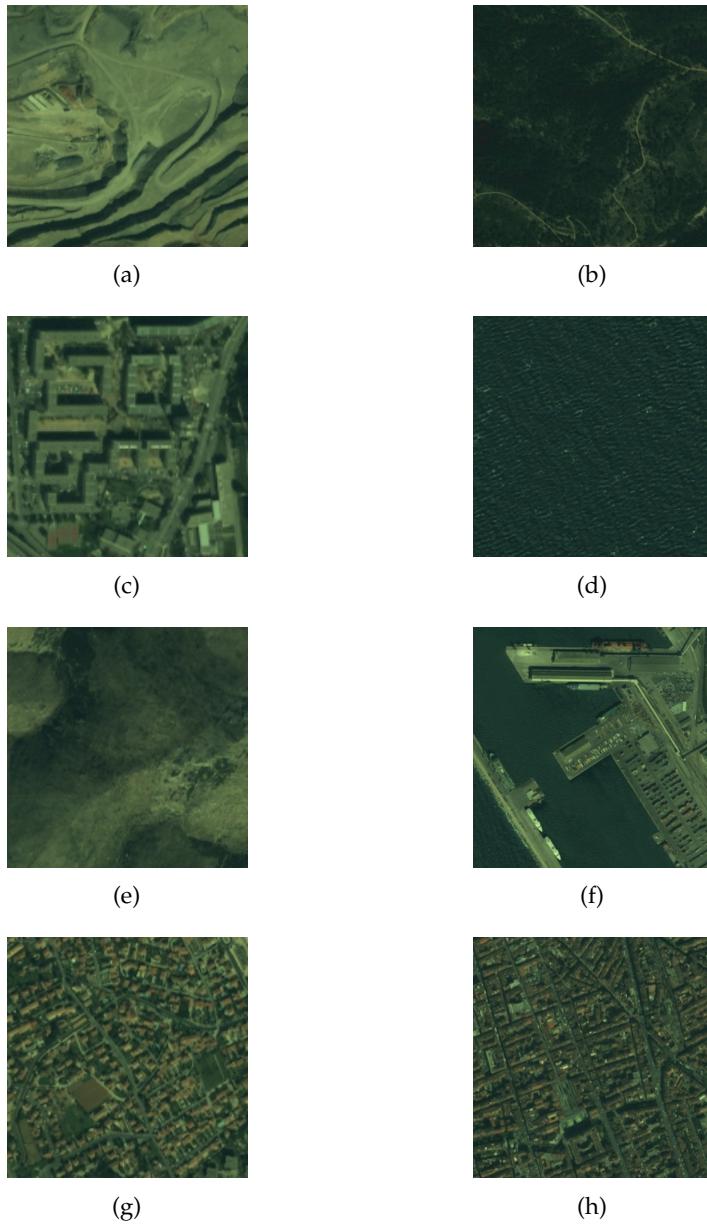


FIG. 7.9 – Exemples d’images de la base d’apprentissage Quickbird de Marseille. (a) : carrières (CA), (b) : espaces de végétation (EV), (c) : grands bâtiments (GB), (d) : mer (ME), (e) : montagne (MT), (f) : port (PO), (g) : zones résidentielles (ZR), et (h) : zones urbaines (ZU).

les différentes distributions de topics. Nous avons utilisé ici la distance du cosinus, très populaire dans les systèmes de recherche d’images et de texte, et qui se définit comme suit entre deux vecteurs **a** et **b** :

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} \quad (7.7)$$

Lorsque les distributions des topics sont proches, les classes correspondantes sont

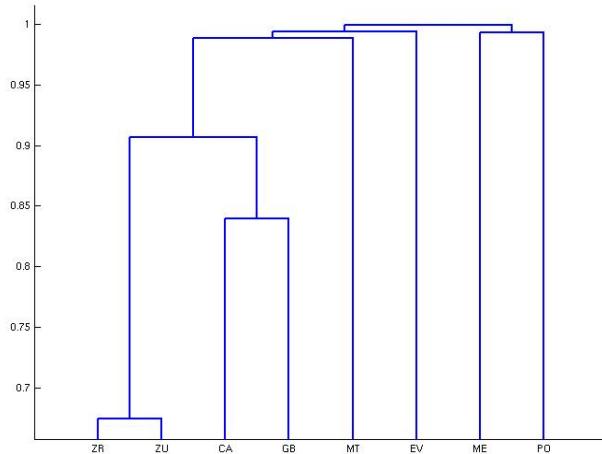


FIG. 7.10 – Cette structure hiérarchique indique les relations entre les 8 classes, basées sur le calcul de la distance du cosinus entre les différentes distributions des topics. Plus ces dernières sont similaires, plus les classes correspondantes sont proches dans la structure hiérarchique.

aussi proches dans la structure hiérarchique. Nous pouvons ainsi prévoir quelques confusions entre les *zones résidentielles* et les *zones urbaines*, entre les *grands bâtiments* et les *carrières*, mais aussi entre les *espaces verts* et les *montagnes*.

L'image de test est complètement annotée, suivant la procédure sus-décrise. De manière globale, les résultats semblent intéressants. Plus rigoureusement, l'évaluation de l'annotation effectuée selon la procédure décrite dans la section 7.2.2.3, donne les résultats consignés dans le tableau 7.2. En effet, ce dernier indique le degré de reconnaissance de chacune des régions de test avec notre approche d'annotation. Ainsi, chaque classe a une à trois régions de test choisies pour chaque concept sémantique.

L'examen du tableau 7.2 montre qu'il est possible de scinder les différentes classes en trois catégories. La première regroupe les classes qui sont très bien identifiées telles que les *grands bâtiments*, la *mer* et les *zones urbaines*. En effet, chacune des régions d'évaluation pour chacune de ces classes est presque parfaitement déterminée (au moins à 98%). La deuxième catégorie rassemble 4 classes qui sont assez bien reconnues. En effet, nous avons obtenu jusqu'à 97% de bonne identification pour les *ports*, 94% pour les *espaces verts*, 90% pour les *zones résidentielles* et 90% pour les *carrières*. Le dernier groupe enfin, se réduit à la classe des *montagnes* qui est mal reconnue, avec seulement 40% de bonne classification.

Les erreurs de reconnaissance présentes sont principalement liées aux similitudes spectrales fortes avec d'autres classes. Les *carrières*, par exemple, se confondent quelquefois aux *grands bâtiments* dont les toits ont généralement une forte intensité lumineuse (figure 7.11). Ce type d'erreur est accentué par le fait que la modélisation des images par l'Allocation Dirichlet Latente implique de travailler de manière locale sur un voisinage de pixels, représenté par des mots. Dans la littérature, un mot est en général une région de l'image issue d'une segmentation, par analogie avec l'analyse de texte. Dans

TAB. 7.2 – Evaluation de l'annotation en utilisant les caractéristiques spectrales uniquement : degrés de reconnaissance de différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.90	0.94	1	1	0.40	0.97	0.90	1
Région 2	-	0.92	0.99	1	-	0.89	0.81	1
Région 3	-	0.65	0.99	-	-	-	0.75	0.98

nos expérimentations, un mot a été considéré comme étant une fenêtre de pixels, obtenue en découplant l'image suivant une grille régulière. Il est bien vrai que la taille du mot a été déterminée de manière empirique, cependant, une fenêtre de pixels ne saurait capturer toute l'information qui pourrait être contenue dans un segment de l'image. Par ailleurs, elle pourrait mélanger des informations contenues dans plusieurs régions, rendant ainsi impure ou biaisée, l'information délivrée par le mot. C'est un biais qui influence de manière non négligeable le calcul des mots visuels, et donc la détermination des topics latents. En effet, pour peu que deux classes aient des similitudes spectrales, ce biais pourrait fortement avantager l'une des classes aux dépens de l'autre. Et pour cette dernière classe, les confusions seront plus ou moins fortes en fonction des mots appartenant initialement aux différentes régions de test. Les remarques faites ci-dessus pour les mots valent également pour les documents à classifier, puisque ces derniers sont obtenus en découplant la grande image à annoter suivant un maillage régulier. Un document peut par conséquent combiner des informations appartenant à plusieurs classes. La procédure de prise en compte de l'information spatiale, décrite dans la section 7.1.2, permet d'atténuer les effets causés par ce mode d'obtention des documents.

Par ailleurs, pour les *espaces verts*, les *zones résidentielles* et les *ports*, certaines régions d'évaluation sont bien identifiées et d'autres un peu moins bien. Par exemple, il existe une différence de taux de reconnaissance de 29% entre les première et troisième régions de test de la classe des *espaces verts*. Cela peut s'expliquer en considérant la nature des régions voisines et adjacentes à la région de test. En effet, si la région de test est entourée de régions appartenant à des classes sémantiques proches de la sienne, la qualité de la reconnaissance peut en pâtir, puisque l'utilisation des fenêtres de pixels pour la modélisation LDA, obtenues par découpage selon une grille régulière, favorise les confusions au niveau des frontières des régions, et a fortiori, celles de régions appartenant à des classes sémantiques proches. La figure 7.13 illustre le cas des *espaces verts*, où l'une des régions de test, principalement entourée de régions appartenant aux *zones résidentielles*, n'est reconnue qu'à 65%, parce que confondue à cette dernière classe au niveau des frontières (figure 7.13(a)), tandis que l'autre zone d'évaluation, principalement adjacente aux *grands bâtiments* avec lesquels les risques de confusions sont moindres, a 94% de bonne identification (figure 7.13(b)). De même, comme le montre la figure 7.11, la zone d'évaluation des *montagnes* est principalement assimilée aux *espaces verts*, car non seulement ces deux classes ont des ressemblances flagrantes d'un point de vue radiométrique, mais en plus la région de test des *montagnes* est voisine d'une région appartenant aux *espaces verts*.

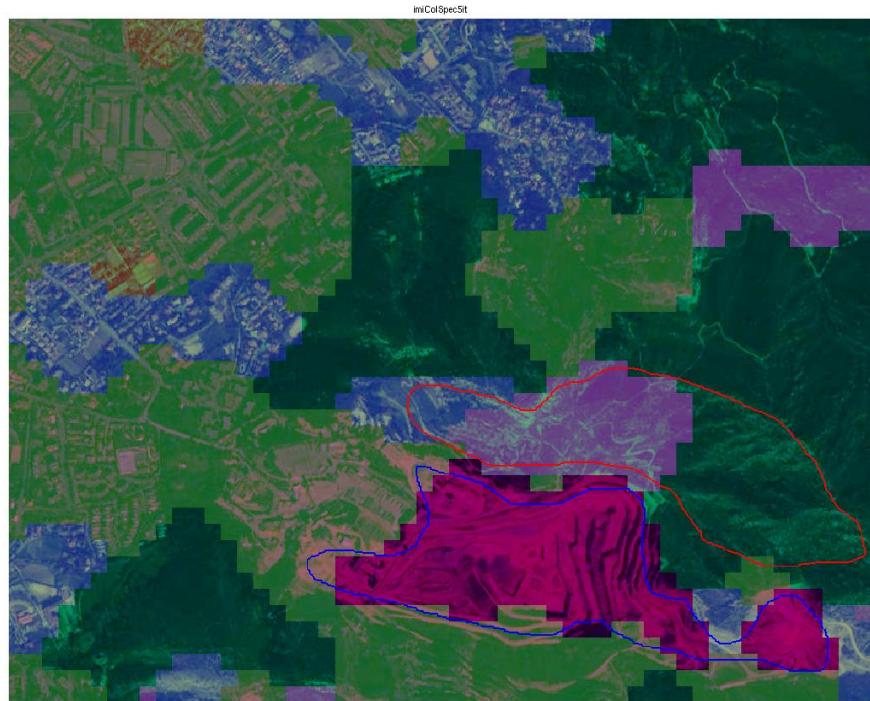


FIG. 7.11 – Régions de test pour les *carrières* (détourée en bleu) et les *montagnes* (détourée en rouge). Les carrières, de forte intensité, sont quelquefois confondues avec les grands bâtiments dont les toits sont lumineux. La montagne, quant à elle, est essentiellement assimilée aux espaces verts.



FIG. 7.12 – Légende associée aux annotations obtenues sur les images Quickbird de Marseille.

Influence de la texture Pour juger de l'apport des primitives de texture basées sur les filtres miroirs en quadrature (QMF), nous avons mené des expérimentations avec les mêmes régions de test. Les résultats obtenus, contenus dans le tableau 7.3, sont comparés à ceux du tableau 7.2.

Nous constatons que, par rapport au cas précédent où seules les caractéristiques spectrales ont été utilisées, il n'y a aucune ou très peu de différences dans les taux de reconnaissance des classes *mer*, *zones urbaines* et *carrières*, les deux premières étant déjà très bien identifiées sans l'utilisation de la texture. L'amélioration est notable pour la classe

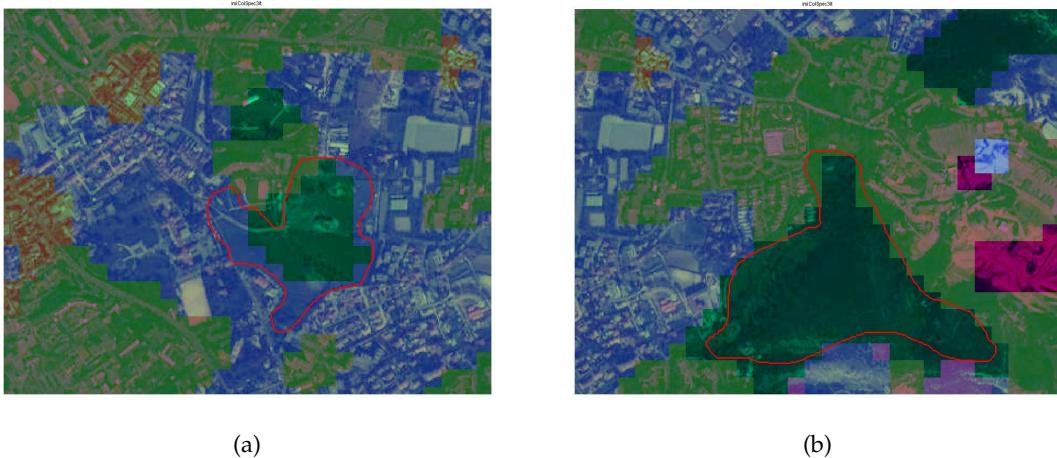


FIG. 7.13 – Exemples de régions de test pour les *espaces verts*. (a) : la région de test est principalement entourée de régions appartenant aux *zones résidentielles* et n'est reconnue qu'à 65%. (b) : la région de test est principalement adjacente aux *grands bâtiments* avec lesquels les risques de confusions sont moindres et a 94% de bonne identification.

TAB. 7.3 – Evaluation de l'annotation en utilisant les caractéristiques spectrales et texturelles : degrés de reconnaissance de différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.89	0.81	0.95	1	0.77	0.65	0.82	1
Région 2	-	0.78	0.97	1	-	0.85	0.78	1
Région 3	-	0.78	0.96	-	-	-	0.99	0.95

des *montagnes*, qui passe de 40% à 77%, mais n'est toujours pas très bien identifiée. Les *grands bâtiments* et les *ports* sont moins bien reconnus, ce qui n'est guère surprenant, les primitives de texture ne caractérisant généralement pas de manière pertinente, ce type de classes haute résolution, riches en structures linéaires. En ce qui concerne les *espaces verts* et les *zones résidentielles*, le taux de reconnaissance augmente pour certaines régions de test, tandis qu'il décroît pour d'autres. Cela est probablement lié aux confusions entre les différentes classes, accentuées par voisinage de la région de test. La figure 7.14 montre l'exemple d'une région d'évaluation des *zones résidentielles*, dont l'ajout de la texture a amélioré le taux de reconnaissance.

De manière générale, l'apport de la texture est moindre, et pour la majorité des classes, la qualité de l'identification est même dégradée. Nous pensons que ceci est en grande partie dû à la résolution de l'image qui est assez haute (2.44 m). En effet, les classes qui ont un niveau sémantique plus élevé, telles que les *grands bâtiments*, sont plutôt moins bien reconnues. Dans la suite, nous ne considérerons donc pas les caractéristiques texturales et utiliserons uniquement les caractéristiques spectrales pour les expérimentations.

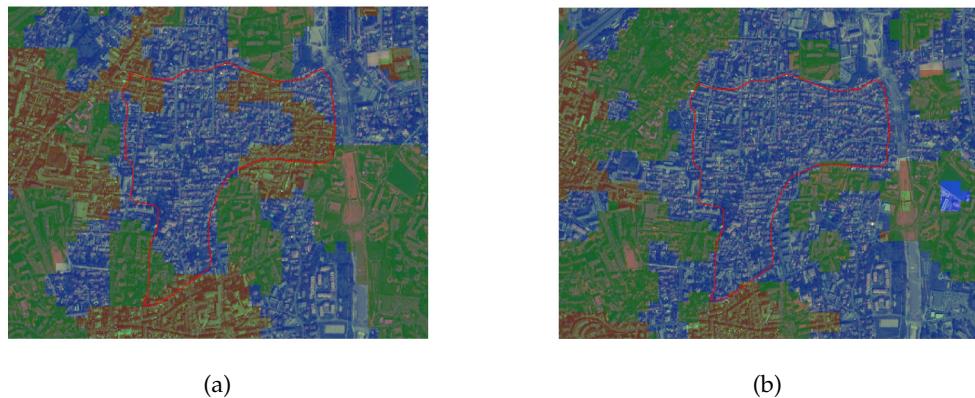


FIG. 7.14 – Exemple d'une région de test de la classe des *zones résidentielles*. L'annotation a été opérée en utilisant les caractéristiques spectrales pour 75% de reconnaissance (a), puis les caractéristiques spectrales et texturelles pour 99% de reconnaissance (b).

7.2.3 Cas de l'assimilation des mots à des régions de l'image

Dans cette section, nous ne considérons plus un mot comme une fenêtre de pixels obtenue par découpage de l'image suivant une grille régulière, mais plutôt comme une région de l'image issue d'une segmentation. En effet, une segmentation judicieuse de l'image produit des mots plus "intuitifs" et plus significatifs, porteurs d'une information moins bruitée.

Nous souhaitons donc annoter les images, en suivant la même approche que précédemment, résumée dans la figure 7.1. Ainsi, les images d'apprentissage, ainsi que l'image de test, sont partitionnées en régions à l'aide de l'algorithme de segmentation Mean shift, qui permet d'éviter la présence de régions de taille inférieure à un seuil donné. Le vocabulaire visuel est obtenu par quantification des primitives spectrales et géométriques extraites des régions des images de l'ensemble d'apprentissage. Puis, l'apprentissage par le LDA est effectué de la même manière que précédemment. En ce qui concerne la grande image de test, la récupération des documents (imagettes) à classifier est plus problématique que dans les expérimentations précédentes. En effet, les documents étaient découpés dans la grande image suivant une grille régulière, ce qui coïncidait avec la définition des mots qui étaient obtenus de la même manière, mais avec une grille plus serrée : un mot n'était donc jamais partagé entre deux documents.

Mais ici, si nous essayons d'obtenir les documents de la même manière, une région ayant une forme quelconque peut être partagée entre plusieurs régions, compliquant ainsi le compte des occurrences des mots visuels pour les tests. Pour remédier à cela, plusieurs pistes ont été explorées.

Regroupement par croissance de régions : Ici, l'idée est d'obtenir les documents en découplant la grande image de test suivant une grille régulière. Les régions réparties dans plusieurs documents sont simplement comptées comme une occurrence du mot visuel associé à la région. Les tests sont donc effectués sur l'ensemble des documents en utilisant les modèles appris avec le LDA. Deux documents adjacents qui partagent une région pouvant appartenir à deux classes différentes, il est donc nécessaire d'effectuer un post-

traitement qui mettrait en évidence l'intérêt de cette approche par régions.

Partant de plusieurs documents initiaux de la grande image de test ayant obtenu une probabilité d'appartenance à leur classe supérieure à un certain seuil, utiliser un algorithme glouton, qui pour un document donné, essaierait de lui associer les régions qui lui sont voisines, en respectant la contrainte que la probabilité d'appartenance du document combiné aux nouvelles régions soit supérieure à un seuil.

Cependant, les régions sont dotées d'une étiquette, et une classe sémantique est une combinaison de ces étiquettes. L'algorithme de croissance de régions, pour être efficace, devrait donc pouvoir associer plusieurs régions à la fois à un document, pour essayer de retrouver la distribution des mots visuels dans la classe sémantique. Mais, nous ne disposons pas d'informations pour éviter de le faire de manière aléatoire. De plus, les images de test qui nous intéressent sont de grande taille (6000×6000 pixels pour l'image de test de Las Vegas et 2400×3040 pour celle de Marseille) : la segmentation de l'image de test de Marseille a donné environs 90000 régions. Avec cette méthode, on peut donc s'attendre à des calculs longs et fastidieux.

Définition des documents comme des groupes de régions connexes : Par ailleurs, pour éviter le problème des régions partagées entre plusieurs documents, ces derniers ont été considérés non plus comme des fenêtres de pixels, mais comme des groupes de régions connexes, où chaque région représente un mot. Cela ne change rien à la représentation par sacs-de-mots puisqu'elle ne tient pas compte des relations spatiales entre les régions dans un document. Ainsi, la grande image de tests est partitionnée en groupes de régions connexes.

Cependant, cette méthode présente un inconvénient certain pour certaines classes de mélange comme les *ports*. Tel que nous avons défini cette classe (voir figure 7.9(f)), ils sont composés d'un peu de zones d'eau (mer) et des installations pour les bateaux. Or, lors de la segmentation, la mer est en général déterminée comme une région, donc il ne sera pas possible d'avoir de petites régions de mer à associer à d'autres régions connexes pour constituer les ports. En effet, on aura d'une part une grande région d'eau correspondant à la *mer*, et à côté de petites régions correspondant aux quais.

L'utilisation des régions de l'image comme des mots est intéressante. Pour des applications de recherche d'images dans de grandes bases, elle peut être plus performante que l'utilisation des fenêtres de pixels. Mais dans notre cas, qui est l'annotation de grandes images, la prise en compte de la continuité entre les documents pose un problème assez délicat, dont la résolution peut nécessiter des calculs pénibles. Dans la suite de notre étude, tout comme dans les expérimentations déjà effectuées dans ce chapitre, nous considérerons donc un mot comme une fenêtre de pixels.

7.3 Etudes comparatives

Dans cette section, nous comparons les résultats de classification obtenus avec le modèle LDA, avec ceux produits par les modèles gaussien, et GMM sur une approche pixel, et le classificateur SVM utilisant une approche par sacs-de-mots.

TAB. 7.4 – Résultats obtenus avec le modèle gaussien, en utilisant les mêmes caractéristiques spectrales que celles du tableau 7.2 : degrés de reconnaissance des différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.87	0.90	0.22	1	0.80	0.36	0.35	0.76
Région 2	-	0.60	0.19	1	-	0.09	0.60	0.50
Région 3	-	0.85	0.10	-	-	-	0.43	0.49

7.3.1 Modèles gaussien, GMM et LDA

Dans cette partie, nous comparons les résultats précédents avec ceux que nous pourrions obtenir en utilisant un modèle gaussien (que nous avons déjà utilisé dans le chapitre 4) ou un mélange de gaussiennes (GMM). Les données utilisées pour étayer notre étude appartiennent à la base d’images Quickbird de Marseille, utilisée dans la section 7.2.2.

Comme précédemment, nous souhaitons annoter de grandes images, via une classification automatique supervisée en S classes associées à des concepts sémantiques, combinée à une prise en compte de l’information spatiale. Contrairement au cas précédent où les images de la base de données sont générées en utilisant le modèle LDA, ici, la classification supervisée s’effectue dans le cadre des modèles gaussien et GMM, et est appliquée, non pas sur chaque pixel de l’image, mais sur des fenêtres de taille 4×4 pixels (qui sont en fait les mots), ceci afin de réduire la taille des données, et donc les temps de calcul. En d’autres termes, les tests sont effectués sur une nouvelle image réduite dans laquelle chaque pixel représente une fenêtre de taille 4×4 pixels de l’image initiale. Avec ces deux modèles, le critère de décision sur l’assignation d’un individu de l’ensemble des données à une classe repose sur la méthode du Maximum a Posteriori (MAP) ou Maximum de Vraisemblance (MV) lorsque la distribution a priori est uniforme. Et dans ce dernier cas, le critère de décision est le même que celui utilisé lors de la classification avec le modèle LDA.

Dans la suite, nous comparons les résultats donnés par les modèles gaussien et GMM avec ceux issus du modèle LDA. Nous reviendrons ensuite sur les aspects théoriques de ces modèles et nous intéresserons aux différentes hypothèses qui y sont associées.

7.3.1.1 Expérimentations

Les résultats obtenus avec le modèle gaussien en utilisant les caractéristiques spectrales sont consignés dans le tableau 7.4.

En comparant les résultats de ce tableau à ceux du tableau 7.2 issus de l’utilisation du modèle LDA, nous remarquons d’emblée que la classe *mer* est tout aussi parfaitement reconnue en utilisant le modèle gaussien. Nous constatons également une meilleure reconnaissance de la classe des *montagnes* avec ce dernier modèle qui a donné 80% de bonne identification, contre 40% obtenus avec le modèle LDA. Les *carrières* quant à elles, sont presque aussi bien reconnues, mais présentent beaucoup de fausses alarmes, notamment avec les *grands bâtiments*. En effet, étant donné que la classification se fait au niveau du pixel, et non au niveau du document, l’information de voisinage qui, dans le cas du

TAB. 7.5 – Résultats obtenus avec le mélange de gaussiennes, en utilisant les mêmes caractéristiques spectrales que celles du tableau 7.2 : degrés de reconnaissance des différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.88	0.83	0.68	1	0.69	0.82	0.64	0.91
Région 2	-	0.70	0.67	1	-	0.57	0.78	0.89
Région 3	-	0.85	0.77	-	-	-	0.69	0.74

LDA, apporte généralement des informations complémentaires pour l'apprentissage de la classe, est inexiste ici ; d'où la forte présence des fausses alarmes. Cependant, cette prise en compte de l'information de voisinage dans le modèle LDA est défavorable aux frontières des régions, où les documents peuvent mêler des informations appartenant potentiellement à plusieurs classes. Tandis qu'en effectuant la classification au niveau du pixel, les chances de mieux délimiter la région de test sont plus fortes. La troisième région d'évaluation des espaces verts en est une illustration. En effet, elle était reconnue à 65% avec le LDA, contre 85% avec le modèle gaussien, les confusions au niveau des contours de la région ayant considérablement diminué.

Les autres classes sont en général, moins bien identifiées avec le modèle gaussien. En particulier les *grands bâtiments* et les *ports*, qui sont pourtant très bien reconnus avec le LDA (taux de reconnaissance de 89% à 100%), sont à peine identifiés sous l'hypothèse gaussienne (36% maximum de bonne reconnaissance). Ces dernières classes s'avèrent être des classes un peu plus complexes, qu'il convient peut-être de modéliser avec une loi de mélange. Nous avons donc logiquement effectué des tests avec le modèle de mélange de gaussiennes, afin de comparer les résultats obtenus, particulièrement pour ces classes complexes, à ceux donnés par le LDA, qui lui-même est un modèle de mélange.

Le tableau 7.5 expose les résultats obtenus avec le mélange de gaussiennes. Comme nous l'espérions, les classes *grands bâtiments*, *ports*, *zones résidentielles* et *zones urbaines* sont beaucoup mieux identifiées que dans le cas du modèle gaussien, mais les taux de reconnaissance restent inférieurs à ceux obtenus avec le modèle LDA (tableau 7.2).

A partir de toutes ces expérimentations et analyses, nous pouvons conclure que :

- la classe *mer* est parfaitement reconnue, quelque soit le modèle,
- les *montagnes* sont mieux identifiées par le modèle gaussien, qui devance le modèle LDA de 3% seulement de reconnaissance. La situation est inversée pour les *carrières*,
- les *espaces verts* sont globalement mieux reconnus par le LDA, mais le modèle gaussien n'est pas à exclure,
- en ce qui concerne les autres classes, elles obtiennent de meilleurs taux de reconnaissance avec le modèle LDA, qui l'emporte sur le modèle GMM en termes de performances.

Pour justifier les résultats que nous avons obtenus, nous nous proposons de revenir sur les aspects théoriques de ces modèles, et de remettre en cause les hypothèses qui ont été faites.

7.3.1.2 Modèle gaussien

Ici, nous nous plaçons tout d'abord dans une hypothèse d'indépendance des pixels d'une même région, c'est-à-dire que nous ne tenons pas compte de l'éventuelle dépendance locale entre un pixel et ses voisins, et donc, entre les pixels adjacents de deux classes différentes. Cette hypothèse forte a ses limites en classification automatique car elle conduit généralement à des effets "poivre et sel" dans les images classifiées. Nous avons déjà été confrontés à ce problème dans le chapitre 4, lors de l'apprentissage automatique à l'aide des cartes CORINE Land Cover. Nous y avions alors pallié en utilisant des méthodes contextuelles après la classification, telles que les champs de Markov.

Nous faisons de plus, une hypothèse de stationnarité des processus aléatoires. En d'autres termes, une classe C_s , $s = 1, \dots, S$, représentée dans l'image par n_s régions R_i^s , $i = 1, \dots, n_s$ se caractérisera par sa probabilité a priori p_s constante dans toute l'image, et aura pour chaque région, la même probabilité conditionnelle de classe $p(x | s)$, x étant toute mesure faite en un pixel de l'image.

Sous l'hypothèse de stationnarité des distributions, nous supposons de plus que les données pour chaque classe suivent une loi gaussienne. La gaussianité de chaque classe peut être vérifiée à l'aide des critères de normalité comme le tracé des histogrammes, ou des tests statistiques. Pour cette dernière approche, plusieurs tests ont été proposés dans la littérature pour juger de l'adéquation d'un ensemble de données à la loi normale unidimensionnelle [Jarque & Bera, 1987], ou multidimensionnelle [Mardia, 1970; Smith & Jain, 1988]. En effet, pour les données à plusieurs dimensions, les tests de normalité individuellement pour chaque variable ne sont pas suffisants pour déterminer la multinormalité, même si chaque variable a une distribution gaussienne. Un récapitulatif de ces tests mettant en évidence les avantages et inconvénients de chacun d'eux est présenté dans [Srivastava & Mudholkar, 2003].

Parmi les différentes approches statistiques développées pour tester la multinormalité d'un ensemble de données, les tests de Mardia [Mardia, 1970], basés sur les coefficients d'assymétrie (skewness) et d'aplatissement (kurtosis) multidimensionnels, sont très courants dans la littérature. Des travaux menés par Romeu et Ozturk [Romeu & Ozturk, 1993] et Bogdan[Bogdan, 1999] montrent que ces tests de multinormalité sont parmi les meilleurs en termes de performance.

Tests de multinormalité de Mardia Pour les données unidimensionnelles, les tests statistiques de normalité basés sur les coefficients d'assymétrie et d'aplatissement, comme celui de Jarque-Bera [Jarque & Bera, 1987], utilisent les moments d'ordre 3 et 4 d'une variable centrée réduite pour décrire ces grandeurs. Dans le cas multidimensionnel, les statistiques d'assymétrie et d'aplatissement ont été introduites par Mardia et appliquées pour tester la multinormalité d'un ensemble de données [Mardia, 1970].

Soit $X = \{x_1, x_2, \dots, x_N\}$, un échantillon aléatoire de taille N , obtenu à partir d'une distribution de dimension p . La matrice des distances de Mahalanobis $D = (d_{ij})$ s'exprime comme suit :

$$d_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x}) \quad \text{et} \quad d_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \quad (7.8)$$

où $i, j = 1, \dots, N$, \bar{x} étant la moyenne de X et S sa matrice de variance covariance.

Partant de cette matrice, les mesures d'assymétrie et d'aplatissement multidimensionnels, notées respectivement $\beta_{1,p}$ et $\beta_{2,p}$ sont définies comme :

$$\beta_{1,p} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^3 \quad \text{et} \quad \beta_{2,p} = \frac{1}{N} \sum_{i=1}^N d_i^4 \quad (7.9)$$

Pour une loi normale multidimensionnelle, $\beta_{1,p} = 0$ et $\beta_{2,p} = p(p+2)$. Soient les statistiques S et K suivantes :

$$S = \frac{N}{6} \beta_{1,p} \quad \text{et} \quad K = \sqrt{N} \frac{(\beta_{2,p} - p(p+2))}{\sqrt{8p(p+2)}} \quad (7.10)$$

Si les données sont issues d'une distribution multinormale, la statistique S suit asymptotiquement une distribution du χ^2 à $\frac{p(p+1)(p+2)}{6}$ degrés de liberté et la statistique K suit asymptotiquement une loi normale $N(0, 1)$. Le but est de tester l'hypothèse nulle H_0 que les données suivent une distribution multinormale. En pratique, cette hypothèse est rejetée à un certain seuil α si S est supérieur à la valeur critique donnée par la distribution du χ^2 à $\frac{p(p+1)(p+2)}{6}$ degrés de liberté, ou si K est supérieur à la valeur critique donnée par la loi normale centrée réduite.

Application aux données Afin de tester la gaussianité des données d'apprentissage, nous avons dans un premier temps, observé leurs histogrammes pour vérifier s'ils sont en forme de "cloche". Ce critère subjectif, permet cependant d'avoir une idée des éventuelles distributions non gaussiennes. La figure 7.15 montre un exemple d'histogrammes de la bande infrarouge pour quelques classes.

Parmi tous les histogrammes, celui de la classe *mer* (figure 7.15(a)) est le seul qui a une forme très proche de celle de la distribution gaussienne, tandis que celui des *ports* par exemple (figure 7.15(b)), bimodal, s'en éloigne fortement. Nous pouvons en conclure que la modélisation de la classe *mer* par une loi normale est moins abusive que l'utilisation de la même loi pour modéliser les *ports*. Par ailleurs, les formes des histogrammes de certaines classes telles que les *montagnes* (figure 7.15(c)) et les *zones urbaines* (figure 7.15(d)), sans vraiment être similaires à la courbe gaussienne, n'en sont pas complètement différentes. Nous ne pouvons donc rien en déduire.

Dans tous les cas, nous avons besoin d'effectuer des tests complémentaires, plus objectifs, et tenant compte du fait que nos données sont multidimensionnelles. Nous avons donc éprouvé la multinormalité des données d'apprentissage de chaque classe sémantique, à l'aide des tests de Mardia. Au delà du fait que nous souhaitons savoir pour quelles classes l'hypothèse de multinormalité est vérifiée, le but est aussi de faire un ordonnancement des différentes classes, en fonction de la similarité de leur distribution avec la loi multinormale. Ceci afin d'évaluer le "degré d'erreur" causé par l'acceptation de l'hypothèse gaussienne, qui pourrait se ressentir dans la classification. Le tableau 7.6 indique les valeurs des statistiques S et K , ainsi que les valeurs critiques données par les distributions associées, pour un seuil $\alpha = 0.001$ et la décision de rejet ou de non rejet de l'hypothèse nulle.

Nous constatons que l'hypothèse nulle n'est pas rejetée uniquement pour la classe *mer*. En d'autres termes, l'hypothèse que les données de la classe *mer* suivent une loi normale multidimensionnelle n'est pas rejetée au niveau significatif $\alpha = 0.001$. En effet, la statistique S est bien inférieure à la valeur critique pour l'assymétrie, et la valeur absolue de K est inférieure à la valeur critique pour l'aplatissement. Ce résultat corrobore celui des histogrammes.

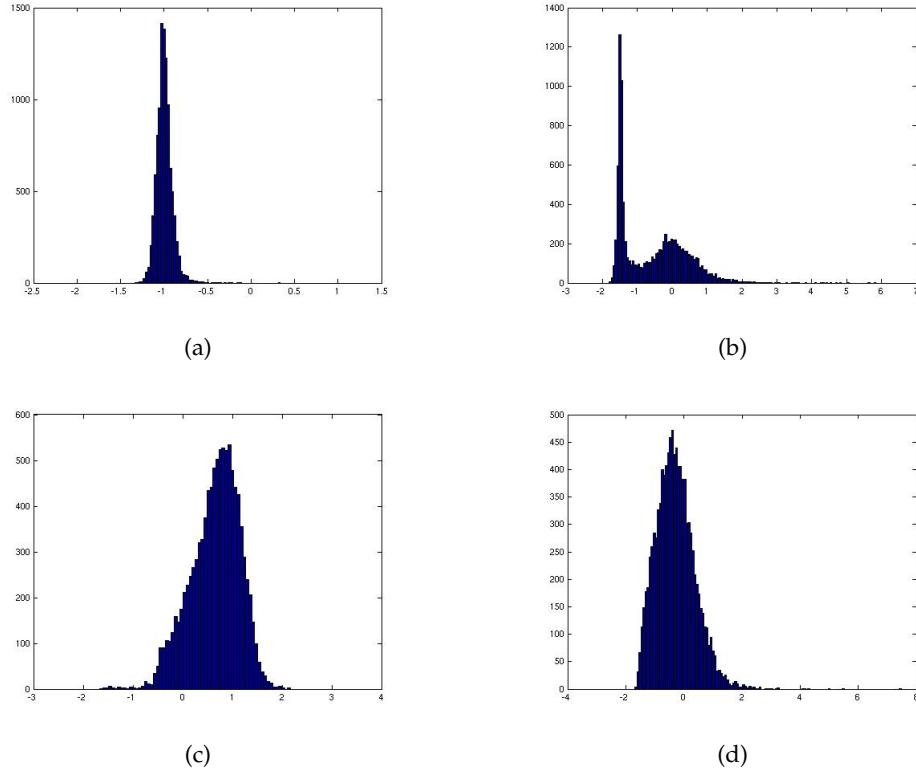


FIG. 7.15 – Histogrammes de la bande infrarouge pour les classes *mer* (a), *ports* (b), *montagnes* (c) et *zones urbaines* (d).

TAB. 7.6 – Résultats des tests d’assymétrie (S) et d’aplatissement (K) de Mardia, avec un seuil $\alpha = 0.001$. L’hypothèse nulle H_0 est rejetée (respectivement pas rejetée) lorsqu’elle est égale à 1 (respectivement 0) au niveau significatif α .

$\alpha = 0.001$, valeur critique assymétrie = 45.31 et valeur critique aplatissement = 3.09								
	CA	EV	GB	ME	MT	PO	ZR	ZU
S	306.01	9.96×10^3	449.46	17.75	108.53	973.85	433.07	7.99×10^3
K	13.38	349.97	9.84	-2.68	0.18	30.68	8.75	211.28
H_0	1	1	1	0	1	1	1	1

En outre, pour les *montagnes*, la statistique $K = 0.18$ est bien inférieure à la valeur critique pour l’aplatissement (3.09), cependant $S = 108.53$ est supérieure à la valeur critique pour l’assymétrie (45.31). L’hypothèse de multinormalité est donc rejetée. Mais, par rapport aux autres classes (*carrières*, *espaces verts*, *grands bâtiments*, *ports*, *zones résidentielles* et *zones urbaines*) pour lesquelles aucune des deux conditions n’est satisfaite, nous pouvons affirmer que la distribution de la classe des *montagnes* est plus proche de la loi multinormale que ne le sont celles des six autres classes.

Ainsi, la vérification de l'adéquation de la distribution de nos données à la loi normale multidimensionnelle montre que la distribution de la classe *mer* est très similaire à celle de la gaussienne, viennent ensuite celle des *montagnes*, et des autres classes.

7.3.1.3 Modèles de mélange

Comme décrit dans la section 3.4.2.3, le mélange de gaussiennes utilise une combinaison convexe d'un ensemble de distributions gaussiennes pour modéliser les observations. L'équation 3.23 exprime la vraisemblance d'un ensemble de données, modélisé par un mélange de K gaussiennes. De manière complète, elle se définit comme suit pour un ensemble X de M données D -dimensionnelles, à partir des équations 3.23 et 3.24 :

$$p(X | \Theta) = \prod_{d=1}^M \sum_{k=1}^K \alpha_k \left(\frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x}_d - \mu_k) \Sigma_k^{-1} (\mathbf{x}_d - \mu_k)' \right) \right) \quad (7.11)$$

où le paramètre global du mélange $\Theta = \{\alpha_k, \mu_k, \Sigma_k, k = 1, \dots, K\}$.

Le modèle LDA, quant à lui, modélise les documents d'une collection par un mélange fini sur un ensemble de topics latents (z_k), eux-mêmes étant modélisés par un mélange sur un ensemble de mots (w_n). Soit D , une collection de M documents \mathbf{w}_d , $d = 1 \dots M$, la vraisemblance du corpus généré par un tel modèle est donnée par :

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d \quad (7.12)$$

obtenue en combinant les équations 5.4 et 5.3.

La vraisemblance du corpus peut s'exprimer sous cette forme grâce à l'hypothèse de "sacs-de-mots" qui stipule que l'ordre des mots dans un document peut être négligé (hypothèse d'échangeabilité des mots dans un document [Aldous, 1985]), de même que l'ordre des documents dans un corpus. En outre, le théorème de De Finetti (1930) établit que toute collection de variables aléatoires échangeables a une représentation sous la forme d'un mélange de distributions. Précisons, par ailleurs, que l'hypothèse d'échangeabilité n'est pas équivalente à l'hypothèse i.i.d (indépendants et identiquement distribués), en raison du paramètre latent de la distribution de probabilité [Blei et al., 2003b]. En effet, si $X_1, X_2, \dots, X_n, \dots$ sont des variables aléatoires indépendantes et identiquement distribuées, alors elles sont échangeables. En revanche, si ces variables aléatoires sont échangeables, alors elles sont indépendantes et identiquement distribuées, conditionnellement au paramètre latent.

L'analyse de ce qui précède indique que le LDA est un modèle de mélange bayésien tout comme le mélange de gaussiennes. Cependant, outre le fait que les distributions composantes sont multinomiales dans le LDA alors qu'elles sont gaussiennes dans le GMM, ces deux modèles sont différents sur bien d'autres aspects.

Modèle hiérarchique L'une des principales différences entre ces deux modèles réside au niveau des proportions de mélange. L'équation 3.23 montre que les proportions de

mélange dans le cas du GMM sont les mêmes pour toutes les données de l'ensemble (α_k fixé, $\forall n$), ce qui n'est pas toujours valide dans le processus de modélisation des données. Le modèle LDA, quant à lui, permet d'avoir des poids de mélange spécifiques à chaque élément de la collection, lui permettant ainsi de mieux s'ajuster aux données d'apprentissage que le mélange de gaussiennes.

En effet, contrairement au modèle GMM non hiérarchique, le LDA est un modèle hiérarchique à trois niveaux, le niveau intermédiaire étant représenté par les topics latents obtenus de manière non supervisée. Ces variables cachées capturent l'information contenue dans les mots, à un niveau "sémantique" un peu plus élevé. En fait, les topics latents doivent correspondre à des catégories d'objets dans les images. Ainsi, les images qui possèdent plusieurs objets sont représentées par une densité mélangée de topics, donnant les proportions de chaque objet dans l'image. De plus, ces proportions des topics, spécifiques à chaque donnée, sont générées à partir d'une distribution Dirichlet (conjuguée de la distribution multinomiale) commune de paramètre α , de telle sorte que les poids des topics pour les différents éléments d'une même collection aient un lien, au lieu d'être choisis indépendamment. Cette propriété permet au modèle LDA d'attribuer une probabilité à des données qui n'appartiennent pas à l'ensemble d'apprentissage, faisant ainsi du LDA, un modèle génératif complet.

Estimation des paramètres Pour le mélange de gaussiennes, $K - 1 + K(D + \frac{D(D+1)}{2})$ paramètres nécessitent d'être estimés, K étant le nombre de composantes du mélange et D , la dimension de l'espace. L'estimation des paramètres se fait par l'algorithme Expectation Maximization (EM) basé sur le maximum de vraisemblance, et a une complexité en $O(KMD^2)$ pour les M exemples d'apprentissage. En ce qui concerne le modèle LDA, le nombre de paramètres à estimer est $K + KV$, où K est le nombre de topics et V la taille du vocabulaire. L'estimation des paramètres se fait également par l'algorithme EM, avec la difficulté que l'étape E ne peut être calculée directement et doit être approximée. En effet, l'inférence exacte n'est en général pas traitable dans le modèle LDA. La solution consiste alors à utiliser des algorithmes d'approximation assez complexes et coûteux tels que l'inférence variationnelle pour l'estimation des paramètres α et β . La procédure d'inférence variationnelle a une complexité en $O(KMV)$ pour l'ensemble des M documents de la collection. Dans nos expérimentations, la taille du vocabulaire est généralement beaucoup plus grande que la dimension de l'espace, donc l'estimation des paramètres du modèle LDA sera en général plus coûteuse. Pour la classe *mer* par exemple, à nombre d'exemples fixé, le temps de calcul pour l'estimation des paramètres est de 4.05×10^{-3} secondes pour le modèle GMM, contre 0.27 secondes pour le modèle LDA, soit environ 65 fois plus. Cependant, pour avoir des performances comparables à celles du LDA, le modèle GMM nécessite d'avoir beaucoup plus d'exemples pour l'apprentissage. Sous les conditions dans lesquelles nous avons effectué nos tests, l'apprentissage de la classe *mer* avec le modèle LDA a nécessité 12 fois plus de temps que l'apprentissage avec le modèle GMM.

Taille de l'ensemble d'apprentissage Pour le mélange de gaussiennes, l'apprentissage est opéré sur des pixels, tandis que pour le modèle LDA, il nécessite des images comme exemples. Les données pour les différents modèles n'étant pas du même type, comparer les tailles des ensembles d'apprentissage paraît insensé. Cependant, une image contenant plusieurs pixels (dans nos expérimentations sur les images Quickbird de Marseille, une

TAB. 7.7 – Nombre de paramètres à estimer et complexité algorithmique des modèles LDA et GMM.

	modèle GMM	modèle LDA
Nombre de paramètres à estimer	$K - 1 + K(D + \frac{D(D+1)}{2})$	$K + KV$
Complexité algorithmique	$O(KMD^2)$	$O(KMV)$

image d'apprentissage pour le LDA contient 64 mots exemples pour le GMM), il sera toujours nécessaire de disposer d'une plus large surface d'images pour la modélisation LDA. Toutefois, sans tenir compte des natures différentes des données, le modèle LDA ne nécessite pas autant d'exemples pour l'apprentissage que le mélange de gaussiennes. En effet, Les résultats donnés par le LDA (tableau 7.2) ont été obtenus avec un ensemble d'apprentissage de 40 images pour chaque classe. Et les performances sont globalement, déjà meilleures que celles du modèle GMM dont l'apprentissage a été fait sur 2560 exemples (tableau 7.5), alors, à plus forte raison si l'apprentissage du GMM avait été fait avec 40 pixels par classe.

7.3.1.4 Sélection de modèles

Tenant compte des observations précédentes, la procédure d'annotation peut être optimisée en introduisant en amont, une étape de sélection du meilleur modèle pour chaque classe. Chacun des modèles est ainsi appris pour chaque classe, puis les performances des différents apprentissages sont évaluées par une procédure de validation croisée, telle que décrite dans la section 3.6.1. La moyenne et l'écart-type de l'erreur des tests sont utilisés pour décider du meilleur modèle pour chaque classe. En supposant que les moyenne et écart-type de l'erreur de test sont les paramètres d'une distribution gaussienne, nous utilisons la divergence de Kullback-Leibler pour mesurer la dissimilarité entre les différents modèles pour une même classe. Soient μ_1, σ_1, μ_2 et σ_2 , les moyennes et écarts-types des erreurs de test de deux modèles \mathcal{M}_1 et \mathcal{M}_2 respectivement, la divergence de Kullback-Leibler dans ce cas [Schowengerdt, 1997] s'exprime par :

$$KL = \frac{1}{2} ((\sigma_1 - \sigma_2)(\sigma_1^{-1} - \sigma_2^{-1})) + \frac{1}{2} ((\mu_1 - \mu_2)^2(\sigma_1^{-1} + \sigma_2^{-1})) \quad (7.13)$$

Lorsque la divergence est inférieure à un certain seuil, on choisit le modèle le plus simple, sinon, le modèle sélectionné est celui qui minimise la moyenne de l'erreur.

Cependant, la nature différente des données en entrée des différents modèles (pixels pour les modèles gaussien et GMM, et documents pour le LDA) rend difficile l'exploitation de cette étape de sélection de modèles. Il serait donc intéressant d'effectuer en outre un parallèle avec un algorithme pouvant s'appliquer sur les documents et utilisant la représentation en mots visuels des images.

7.3.2 Classifications basées sur le LDA et le SVM

Dans cette partie, nous effectuons des tests avec le classificateur SVM (Séparateur à Vaste Marge ou *Support Vector Machines*), très utilisé en analyse de textes pour les tâches

de classification et de recherche, car il a prouvé son efficacité par rapport à d'autres classificateurs (k-plus proches voisins, réseaux de neurones, etc) [Joachims, 1998; Yang & Liu, 1999]. Ici, nous procédon à deux types d'expérimentations. Dans un premier temps, les tests avec le classificateur SVM sur nos images sont effectués comme dans l'analyse de textes, les documents et les mots visuels pour la représentation en sacs-de-mots étant les mêmes que ceux utilisés dans l'analyse avec le modèle LDA. Ensuite, nous opérons également une classification SVM basée sur le pixel, comme nous l'avons fait avec les modèles gaussien et GMM, de manière à mieux juger de l'intérêt d'une représentation en mots visuels pour le classificateur SVM.

7.3.2.1 Classification SVM basée sur une représentation en sacs-de-mots

L'algorithme est appliqué sur une représentation en sacs-de-mots des documents, chaque mot de l'ensemble d'apprentissage étant considéré comme une primitive séparée. En général, la valeur numérique d'un attribut pour un mot (souvent appelé terme) donné est représenté par la fréquence du terme dans le document multiplié par la fréquence inverse du document dans le corpus. Il s'agit de la très populaire représentation tf-idf (*term frequency - inverse document frequency*), qui est une mesure statistique permettant d'évaluer l'importance d'un mot par rapport à un document extrait d'une collection.

Représentation tf-idf : Soient un document \mathbf{w}_d appartenant à un corpus X de M documents, et un terme t_v appartenant à un vocabulaire de taille V , la fréquence du terme dans le document est :

$$tf_{v,d} = \frac{n_{v,d}}{\sum_k n_{k,d}} \quad (7.14)$$

où $n_{v,d}$ est le nombre d'occurrences du terme t_v dans le document \mathbf{w}_d , et le dénominateur est le nombre d'occurrences de tous les termes dans le document \mathbf{w}_d . La fréquence inverse du document est définie par l'expression :

$$idf_v = \log \frac{M}{|\{\mathbf{w}_d : t_v \in \mathbf{w}_d\}|} \quad (7.15)$$

Le dénominateur représente le nombre de documents dans lequel le terme t_v apparaît. Finalement, le poids tf-idf est donné par :

$$tf - idf_{v,d} = tf_{v,d} \times idf_v \quad (7.16)$$

Ainsi, à partir de la représentation des images en mots visuels que nous avons adoptée, il est donc possible d'utiliser le classificateur SVM sur chaque image, munie de sa représentation tf-idf.

Tests et observations : Le vecteur d'occurrences des mots visuels pour chaque image est le même que celui utilisé dans les tests du LDA. Les tests effectués avec le SVM en utilisant la représentation tf-idf obtenue à partir des mots visuels, donnent globalement des résultats intéressants. Cependant, par rapport à l'annotation avec le modèle LDA, beaucoup plus de fausses alertes sont présentes, principalement entre les carrières et les grands bâtiments, et entre les espaces verts et les montagnes. La figure 7.16 montre les annotations d'une même partie de l'image, par le LDA (figure 7.16(a)) et le SVM (figure

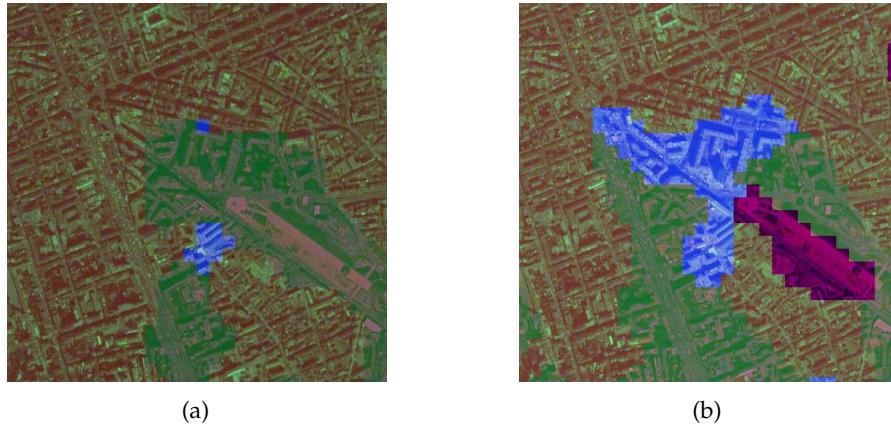


FIG. 7.16 – Annotations d’une portion de l’image par le LDA (a) et le SVM (b). Des confusions plus marquées sont présentes dans la deuxième image, notamment entre les *grands bâtiments* et les *ports*.

TAB. 7.8 – Résultats obtenus avec le SVM basé sur une représentation en sacs-de-mots, en utilisant pour chaque image, les mêmes mots visuels que ceux du tableau 7.2 : degrés de reconnaissance des différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.98	0.93	1	1	0.71	0.79	0.89	1
Région 2	-	0.73	0.99	1	-	0.90	0.78	0.99
Région 3	-	0.80	0.98	-	-	-	0.80	0.91

7.16(b)) respectivement. On note que le SVM présente de plus fortes confusions entre les *grands bâtiments* et les *ports*, mais aussi entre les *grands bâtiments* et les *carrières*.

Les degrés de bonne identification des différentes régions d’évaluation sont consignés dans le tableau 7.8. Nous remarquons que, mis à part les *carrières* et les *montagnes* qui sont mieux reconnues, les six autres classes sont légèrement moins bien identifiées avec le SVM, qu’avec le LDA. La figure 7.17 montre les régions d’évaluation des *montagnes* et des *carrières*. Par rapport à la figure 7.11 représentant l’annotation de la même portion de l’image basée sur le modèle LDA, on voit aisément l’amélioration. Cependant, on note aussi les fausses alarmes qui ont fait leur apparition dans la figure 7.17, notamment, la région des *espaces verts* en bas de l’image est partiellement annotée comme étant des *montagnes*, et quelques petites régions isolées.

Mis à part quelques fausses alarmes, le classificateur SVM a globalement des performances comparables à celles du LDA. En termes de complexité algorithmique, celle du SVM est polynomiale en M , et est comprise entre DM^2 et DM^3 , M étant le nombre de données d’apprentissage et D , leur dimension. Tandis que le LDA a une complexité en $O(KMV)$, où K est le nombre de topics et V , la taille du vocabulaire. Pour un très grand

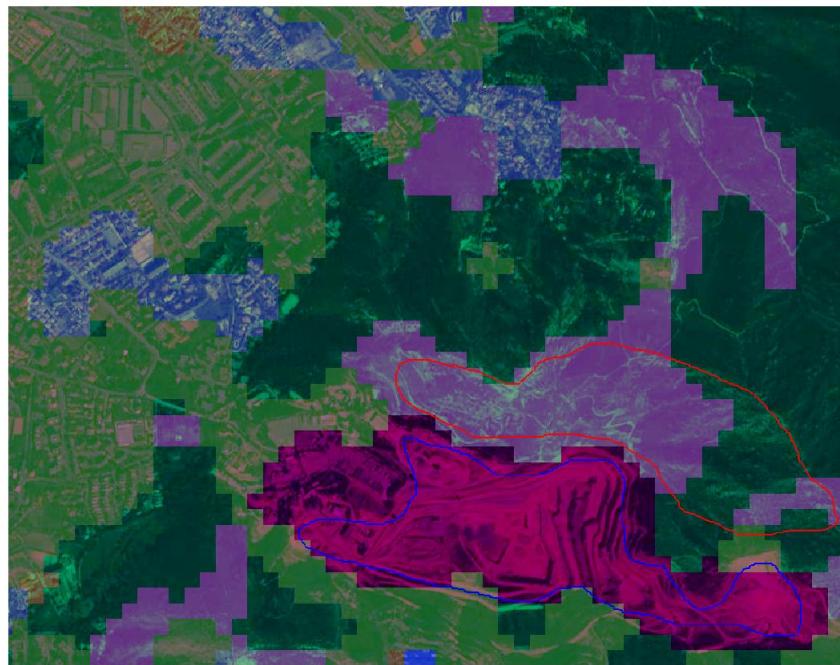


FIG. 7.17 – Régions de test pour les *carrières* (détourée en bleu) et les *montagnes* (détourée en rouge), annotées en utilisant le classificateur SVM.

nombre de données d'apprentissage, le temps de calcul explose avec le SVM. Un autre avantage de la classification basée sur le LDA par rapport au SVM est la notion de topics dans la hiérarchie du LDA, qui, en véhiculant déjà une information sémantique comparé au vecteur des occurrences des mots visuels, permet au LDA de mieux capturer la sémantique de l'image.

7.3.2.2 Classification SVM basée sur le pixel

Dans cette partie, nous souhaitons effectuer la classification SVM sur chaque pixel de l'image, indépendamment des autres. Cependant, étant donné le nombre important de pixels dans notre image de test, la classification est plutôt opérée sur des fenêtres de pixels de taille 4×4 , comme dans la section 7.3.1. Le classificateur SVM utilisé est doté d'un noyau gaussien dont le paramètre choisi est celui qui donne les meilleures performances de validation croisée. Les résultats obtenus sur les zones d'évaluation sont consignés dans le tableau 7.9.

Il en ressort que si cette méthode est plus efficace que la classification bayésienne avec un modèle gaussien, particulièrement pour les classes de mélange (voir tableau 7.4), elle est cependant moins performante que la classification par SVM basée sur les techniques de l'analyse de textes (tableau 7.8). En effet, nous remarquons que mis à part la classe *mer* qui est toujours parfaitement reconnue, toutes les autres classes sont moins bien identifiées en appliquant le classificateur SVM sur le pixel, qu'en utilisant le SVM sur les images munies de leur représentation basée sur les mots visuels, testé dans la section précédente.

TAB. 7.9 – Résultats obtenus avec le SVM, appliqué sur les pixels de l'image individuellement. Le vecteur d'attributs pour chaque pixel est le même que celui du tableau 7.4 : degrés de reconnaissance des différentes régions de test déterminées manuellement pour chaque classe.

	CA	EV	GB	ME	MT	PO	ZR	ZU
Région 1	0.84	0.86	0.69	1	0.57	0.77	0.44	0.84
Région 2	-	0.63	0.70	1	-	0.65	0.51	0.81
Région 3	-	0.75	0.76	-	-	-	0.54	0.73

7.4 Conclusions

Dans ce chapitre, nous avons proposé une approche d'annotation de grandes images satellitaires faisant usage d'un modèle d'analyse statistique des textes : l'allocation Dirichlet latente. Ce modèle hiérarchique utilise une représentation des images en sacs-de-mots, qui néglige les relations spatiales entre les mots visuels. L'approche proposée combine donc une classification supervisée d'imagettes extraites de la grande image à annoter, et une phase de prise en compte de l'information spatiale, qui s'est avérée importante pour améliorer les performances de l'annotation. Cette méthode d'annotation a montré de bonnes performances pour l'identification de classes simples, mais aussi des classes de mélange pour lesquelles son efficacité est largement supérieure à celles des approches basées sur le pixel par exemple, et plus intéressante que celle d'une classification par l'algorithme SVM utilisant une représentation en sacs-de-mots.

Chapitre 8

Conclusions et perspectives

La reconnaissance de la couverture du sol à partir de classifications automatiques est l'une des recherches méthodologiques importantes en télédétection. Par exemple, les cartes CORINE Land Cover, utilisées à l'échelle européenne, sont générées à partir d'images satellitaires qui sont interprétées visuellement par un expert, s'aidant de données exogènes (photographies aériennes, cartes topographiques et thématiques, etc). Des centaines de milliers de km^2 devant être décrits, le processus est long et délicat. Il est donc intéressant de disposer de méthodes automatiques pour l'analyse et l'interprétation de telles bases d'images. Cette thèse s'est intéressée à cette tâche, et visait l'élaboration de méthodes automatiques capables d'apprendre une taxinomie définie par des experts de la production des cartes d'occupation des sols, et d'annoter automatiquement de nouvelles images à l'aide de cette taxinomie.

8.1 Conclusions

Suite à notre analyse de l'état de l'art, nous avons tout d'abord testé l'apprentissage automatique pour la cartographie, via l'usage d'une méthode classique basée sur le maximum de vraisemblance, très présente dans la littérature. Nous l'avons illustré à l'aide de la classification hiérarchique CORINE Land Cover (CLC). En effet, nous avons utilisé les cartes CLC et les images satellitaires multispectrales ayant contribué à la constitution de ces cartes, afin d'apprendre les règles de décision pour la classification de terrains inconnus à partir d'images de télédétection. Nous en avons tiré plusieurs enseignements, en particulier, une classification au niveau du pixel à l'aide de primitives radiométriques et texturales spécifiques est suffisante pour identifier les classes simples telles que les *terres arables* ou les *vignobles*, mais peine cependant à retrouver les classes dites de mélange telles que les *forêts mélangées* ou encore les *réseaux de cultures variées*.

Nous nous sommes alors intéressés plus en profondeur au problème réel qu'est la difficulté d'identifier les classes de mélange avec de telles approches. Nous avons ainsi remis en cause l'étude basée sur une approche au niveau du pixel et avons proposé de nous appuyer sur une approche par régions ou objets pour déterminer les classes de mélange dans les images satellitaires. Cependant, les approches par régions de la littérature classifient chaque région individuellement à partir des primitives extraites. Nous avons donc choisi de représenter les images sous une forme particulière basée sur les régions, de manière à profiter non seulement des propriétés de chacune d'elles, mais aussi des relations existant entre elles. En ce qui nous concerne, l'intérêt de cette représentation, dite en

mots visuels, est qu'elle favorise l'identification des structures de l'image, tout en permettant d'exploiter des outils de l'analyse de textes qui ont montré leur efficacité dans le domaine de la fouille de données textuelles et en classification d'images multimédia. Ainsi, considérant qu'une structure de l'image, qui est un groupe de mots visuels (régions) connexes, appartient à une classe de mélange, nous avons proposé de tirer profit de la compositionnalité sémantique et des modèles probabilistes de l'analyse de textes, et de les appliquer aux images satellitaires, ceci afin d'améliorer l'identification des classes de mélange.

Tout d'abord, la représentation en mots visuels a été utilisée afin d'appliquer le principe de compositionnalité sémantique aux images satellitaires. En effet, nous avons mis en valeur les relations spatiales et plus précisément d'adjacence entre les différents mots visuels. L'importance de la prise en compte de ces relations a été démontrée de deux manières : tout d'abord via une approche non supervisée, illustrée par la détection des bâtiments et des ombres qui leur sont associées, mais aussi une approche supervisée, appliquée pour identifier la classe des *forêts mélangées* dans CLC. Pour ce dernier exemple, nous avons montré que, par rapport à la classification basée sur le pixel où quasiment aucune région de forêt mélangée n'est détectée, l'amélioration est notable. La méthode proposée semble robuste à la résolution des images, les tests ayant été effectués sur des images à faible résolution (SPOT à 20 m) et à très haute résolution (Pelican à 50 cm).

Par ailleurs, nous avons utilisé les techniques d'analyse statistique de textes pour annoter les images satellitaires et en particulier, identifier les classes complexes telles que celles de mélange. Ces méthodes sont basées sur une représentation dite en "sacs-de-mots" qui ne tient pas compte des relations spatiales entre les mots visuels. Nous avons ainsi proposé une approche d'annotation sémantique de grandes images à l'aide de concepts sémantiques définis par l'utilisateur. Cette approche combine une étape de classification supervisée des imagettes extraites de la grande image à classifier, et une phase de prise en compte de l'information spatiale entre les imagettes. Evaluée sur des régions de test issues d'images Quickbird, cette méthode a donné de très bonnes performances pour la classification en général, mais surtout pour l'identification des classes de mélanges telles que les *ports* ou les *banlieues résidentielles*. En effet, les études comparatives que nous avons menées ont montré que l'approche par sacs-de-mots utilisant le modèle textuel LDA est plus efficace sur les classes complexes que les modèles gaussien et mélange de gaussiennes qui utilisent une approche pixel classique. En outre, comparé à l'algorithme SVM utilisant une représentation des images en sacs-de-mots, l'approche que nous avons proposée présente des avantages intéressants, entre autres, la réduction des fausses alarmes.

8.2 Perspectives

L'analyse que nous avons menée ainsi que les résultats obtenus ouvrent un certain nombre de perspectives pour des travaux futurs, à court et à long terme.

Robustesse de l'approche d'annotation proposée A court terme, des traitements supplémentaires peuvent être mis en oeuvre, afin d'améliorer la robustesse de la méthode d'annotation proposée. Par exemple, une classe de rejet pourrait être introduite, pour les cas où les classes sémantiques définies par l'utilisateur ne couvrent pas la totalité de l'image, ou alors lorsqu'il existe une ambiguïté sur la classe d'un échantillon de test

et qu'aucune décision ne peut être prise de manière sûre. Cette classe de rejet peut être définie en précisant un seuil sur la probabilité d'appartenance d'un échantillon à chacune des autres classes. Si aucune des probabilités n'est supérieure à ce seuil, l'échantillon est alors attribué à la classe de rejet. Ce seuil peut être déterminé de manière automatique, ou alors fixé par un expert de manière à ajuster les résultats.

Par ailleurs, la méthode d'annotation sémantique que nous avons proposée a été évaluée sur des images Quickbird à haute et très haute résolution. Il pourrait également être intéressant de l'évaluer sur des images à plus faible résolution, afin de juger de la robustesse de la méthode à la résolution des images.

Types de mots et de documents Dans nos travaux, nous avons plus souvent considéré qu'un mot est une fenêtre de pixels. Nous avons également mené quelques expérimentations avec des mots comme des segments de l'image, sans résultats concluants à cause de la complexité des traitements à effectuer. Une alternative pourrait être de considérer qu'un document est une région de l'image, obtenue après une segmentation grossière. Dans ce cas, les mots seraient des pixels de la région. La procédure serait alors exactement la même, sauf que on classifierait plutôt les régions de l'image. Pour essayer de pallier l'effet de la segmentation grossière, on pourrait exploiter la multirésolution, ou alors effectuer plusieurs segmentations avec des algorithmes différents, et/ou en faisant varier leurs paramètres, et classifier séparément chaque image segmentée. Puis, afin de déterminer la meilleure annotation de l'image, une sorte de consensus, tenant compte de l'appartenance d'un pixel à une même classe dans les différentes classifications, pourrait être bénéfique.

Introduction de la connaissance externe La difficulté d'apprendre les cartes CORINE Land Cover nous conduit à nous interroger sur l'introduction de la connaissance externe pour améliorer le traitement de la sémantique. En effet dans notre étude, nous avons pu estimer la quantité d'information que la seule image satellitaire apporte à la constitution de la carte CLC. Nous avons ainsi constaté l'importance des informations exogènes (cartes topographiques IGN, photographies aériennes, etc) utilisées par le photointerprète, qui pourraient sensiblement accroître la qualité de l'identification des classes de mélange en particulier. Mais alors, se pose le problème délicat de la représentation de cette connaissance externe. Avec une carte topographique par exemple, une approche simple pourrait consister à utiliser la légende associée à la carte, qui est donc annotée. Cette dernière est fusionnée à la carte CLC, en gardant à chaque fois la plus petite région. Chaque région de la nouvelle carte est associée à un couple constitué de sa classe dans la carte topographique et sa classe dans CLC. Les différents couples constituant un ensemble de sous-classes, l'apprentissage est effectué pour chaque couple possible, à partir de zones d'apprentissage de l'image satellitaire. Chaque région de l'image de test sera donc attribuée à une sous-classe, et récupérera la classe CLC correspondante. Cette démarche, qui n'est pas optimale, pourrait permettre de constituer une hiérarchie entre les classes CLC et celles de la carte topographique.

Combinaison de modèles Nous avons montré dans le dernier chapitre de ce document, que pour l'annotation des images satellitaires, certains modèles décrivaient mieux certaines classes. Une direction de recherche, que nous n'avons pu explorer dans le temps imparti, est de combiner l'utilisation de ces différents modèles afin d'améliorer l'identifi-

cation de chaque classe. Une étape de sélection de modèles serait alors nécessaire, pour choisir pour chaque classe, le modèle qui la représente le mieux. Il faudrait tenir compte des types de données sur lesquelles s'appliquent les modèles (pixels, documents, etc), mais également du coût, en termes de temps de calcul, que pourrait nécessiter une telle opération d'optimisation.

Relations spatiales En ce qui concerne les relations spatiales, dans cette thèse nous nous sommes restreints aux relations d'adjacence entre les mots visuels, mais l'importance des autres types de relations n'est pas à démontrer. En effet, afin de reconnaître les structures dans les images, il serait également possible d'utiliser les relations spatiales métriques pour tenir compte de la distance séparant les régions d'intérêt, ou les relations directionnelles qui font appel à une direction de l'espace ou du plan (à gauche de, à droite de, etc). Pour ces relations qui sont souvent imprécises, il est plus intéressant d'utiliser des représentations floues, en introduisant un degré de satisfaction de la relation entre les deux objets. Ces relations floues peuvent être modélisées par exemple à partir d'outils de morphologie mathématique floue. Ceci rentre dans le domaine du raisonnement spatial. Comme dans le cas de l'adjacence, le degré de satisfaction de la relation métrique ou de direction entre deux objets de l'image pourrait être stocké dans une matrice non symétrique, ou une liste. Cette information pourrait faire émerger une certaine régularité dans l'image, en détectant par exemple des objets ayant la même orientation, ou des couples d'objets équidistants. On obtiendrait ainsi des indications sur la localisation de certaines structures dans l'image.

Par ailleurs, nous ne nous sommes intéressés qu'aux relations binaires entre les régions. Les relations d'ordre supérieur (ternaires par exemple), pourraient accélérer le processus d'identification des structures de l'image : par exemple, une "rue" est entre deux "pâtés de maisons", moyennant cependant un algorithme plus complexe.

Annexes

Annexe A

CORINE Land Cover

CORINE Land Cover est un inventaire biophysique de l'occupation des terres, fournant une information géographique de référence pour 29 Etats européens et pour les bandes côtières du Maroc et de la Tunisie. Cette description repose sur une taxinomie précise afin d'assurer la consistence des résultats produits par les photointerprètes. Cette base de données géographiques est produite, gérée et utilisée à l'aide d'un Système d'Information Géographique (SIG).

A.1 Description

Historique

Débuté en 1985, CORINE Land Cover [Buttner et al., 2004] est la plus grande base de données du programme CORINE¹. Une première version de la base, dite CLC90, a été réalisée à partir d'images satellitaires Landsat MSS et SPOT XS acquises entre 1987 et 1994. Les différents pays européens ont été chargés de réaliser une base de données nationale d'occupation du sol que la Commission Européenne (CE) a ensuite centralisée.

Le besoin d'une base de données mise à jour, exprimé par plusieurs utilisateurs au niveau national et européen, a conduit en 1999 à des travaux préparatoires pour l'année de référence 2000 : le projet "IMAGE & CLC2000". La mise à jour a été achevée en 2004 avec des images Landsat ETM+ acquises en 2000 plus ou moins 1 an (CORINE Land Cover 2000 ou CLC2000). Il était ainsi possible de mettre en évidence les zones où l'occupation du sol a évolué (extension des villes et des forêts, recul des prairies, création d'autoroutes, ...) sur une superficie supérieure à 5 ha.

Pour assurer une couverture totale et maximiser la compatibilité avec l'inventaire précédent, "IMAGE & CLC2000" fait appel aux expertises locales existantes et nécessite l'accès aux données utilisées lors du premier inventaire CORINE Land Cover.

CLC2000 couvre environ 4.5 millions de km^2 dont 550 000 en France, et représente un véritable référentiel d'occupation du sol, proche par la date des recensements de la population (1999) et de l'agriculture (2000).

¹Le programme CORINE : programme de CO-ordination de l'INformation sur l'environnement, a pour objectif, de fournir une information fiable et régulière pour la gestion et l'aide à la prise de décision de la Commission Européenne en matière d'environnement. Ce projet est un effort collectif de différents pays de l'Europe, pour constituer une base de données des couvertures agronomiques, environnementales et d'aménagement du territoire des pays européens.

Principes de base

L'information produite par CORINE Land Cover devant être homogène, strictement comparable pour tous les pays concernés et susceptible d'être mise à jour périodiquement, trois principes fondamentaux [Ins, 2005b] ont été définis afin de satisfaire ces conditions : l'échelle de travail, la définition de la superficie minimale des unités cartographiées et la nomenclature d'occupation du sol.

Echelle de travail L'échelle de travail choisie est 1/100 000 car bien adaptée aux besoins nationaux et européens de suivi et de gestion de l'environnement ou d'aménagement de l'espace. De plus, cette échelle est compatible avec les contraintes de coût de production et d'actualisation ainsi qu'avec celles des délais de réalisation et permet d'envisager une mise à jour régulière.

L'unité spatiale L'unité spatiale au sens de CORINE Land Cover est une zone dont la couverture peut être considérée comme homogène, ou être perçue comme une combinaison de zones élémentaires représentant une structure d'occupation. L'homogénéité d'une zone est évaluée visuellement : au moins 75% de la superficie de l'unité paysagère doit appartenir à une même classe d'occupation du sol.

La surface de la plus petite unité cartographiée (seuil de description) est de 25 hectares au sol, soit par exemple un carré de 5 mm de côté à 1 : 100 000. Ce choix a été fait pour faciliter la digitalisation des documents d'auteur et l'impression de cartes lisibles. Dans la pratique, des éléments de l'occupation du sol inférieurs à cette superficie ont parfois été cartographiés lorsqu'ils avaient une importance significative (village, infrastructure de loisirs, ...). En effet, en raison du morcellement parfois important de ses paysages, la Belgique [Ins, 1995] a reçu l'autorisation de représenter des surfaces pouvant descendre jusqu'à 10 ha lorsqu'elles avaient une importance significative.

De même, la largeur minimale des éléments linéaires représentés est de 100 m au sol, soit 1 mm à 1 : 100 000. Cependant, certains tronçons plus étroits ont été cartographiés de manière à conserver la continuité des éléments linéaires dans la base de données (par exemple, tracé de la Meuse à partir de la frontière française vers la Belgique).

Nomenclature La nomenclature adoptée au niveau européen est conforme aux classes proposées dans le guide technique CORINE Land Cover [Bossard et al., 2000] ainsi que dans [Ins, 2005a]. C'est une classification hiérarchisée en 3 niveaux et 44 postes répartis selon 5 grands types d'occupation du territoire : territoires artificialisés, territoires agricoles, forêts et milieux semi-naturels, zones humides et surfaces en eau. Les différentes classes sont présentées en détail dans la deuxième partie de ce chapitre.

L'élaboration de cette nomenclature d'occupation des sols a été orientée sur l'occupation biophysique du sol et non sur son utilisation [Gregorio & Jansen, 1998] ; elle privilégie donc la nature des objets (forêts, cultures, surfaces en eau, roches affleurantes...) plutôt que leur fonction socio-économique (agriculture, habitat,...). En acceptant les définitions proposées par le guide technique, la légende ne fait pas uniquement référence à l'occupation du sol, mais aussi parfois à son utilisation ("land cover" et "land use"). Dès lors, l'interprète est parfois confronté à des zones pouvant être classées de deux façons différentes. Dans de telles situations, l'occupation du sol a toujours été privilégiée. Par exemple, un dépôt militaire situé en pleine forêt de feuillus peut être classé dans la classe

1.2.1. (*zones industrielles ou commerciales*) en privilégiant l'utilisation du sol ou dans la classe 3.1.1. (*forêt de feuillus*) en privilégiant l'occupation du sol.

L'un des objectifs de l'inventaire CORINE Land Cover est de cartographier l'ensemble des territoires européens sans recourir à un poste "territoire non classé" ou "autre" et, dans un souci de cohérence et d'homogénéité européenne, en définissant le mieux possible chacun des postes de la nomenclature utilisée. Mais pour satisfaire au critère de superficie minimale des unités cartographiées (25 ha), certains modes d'occupation des terres ont dû être regroupés au sein de postes appelés "postes à caractère mixte".

- Les systèmes parcellaires et culturaux complexes (2.4.2) :
Il s'agit ici de petites parcelles de cultures annuelles diversifiées, de prairies et/ou de cultures permanentes. Aucune de ces trois catégories ne répond au seuil de 25 hectares (ni au seuil de tolérance qui est d'environ 15 hectares) et ces terres arables, prairies ou vergers occupent chacun moins de 75% de la superficie totale de l'unité paysagère. Ce poste caractérise donc la diversité locale des modes d'occupation des terres.
- Les territoires principalement occupés par l'agriculture, avec présence de végétation naturelle importante (2.4.3)

Ce poste correspond à des territoires agricoles interrompus par des espaces naturels importants (landes, pelouses, ...). Il est caractérisé par des terres agricoles qui occupent entre 25% et 75% de la surface totale de l'unité paysagère, mais comme pour le poste précédent, aucun sous-ensemble homogène répondant au seuil de description de 25 ha ne peut être isolé.

Par ailleurs, lorsqu'une zone homogène est inférieure à 25 hectares, des règles précises de généralisation ont été adoptées. Ces règles varient suivant la classe concernée et sont détaillées dans [Bossard et al., 2000]. Entre autres, la généralisation du tissu urbain discontinu (classe 1.1.2.) a été limitée à une distance de 300 mètres entre deux maisons le long d'une route, pour garder la caractéristique d'occupation du sol des rues de villages. De même, les pistes des aéroports doivent être entourées d'une zone "tampon" d'un minimum de 100 mètres.

Génération de CORINE Land Cover

La base de données CORINE Land Cover est générée à partir :

- d'imagerie satellitaire (en particulier Landsat Multispectral et SPOT Multispectral, mais aussi AVHRR [Hastings & Tateishi, 1998]) : les images Landsat TM (Thematic Mapper) sont les plus souvent utilisées en raison de 3 caractéristiques : une résolution de 30 m adaptée à l'échelle du 1 : 100 000 choisie, un choix large de bandes spectrales (7 bandes dont 4 dans l'infra-rouge) permettant une bonne discrimination des types végétaux et une grande surface couverte par l'image (185 km de côté).
- de cartes topographiques de l'IGN² : les cartes à 1 : 100 000 de l'IGN ont servi de base géométrique à la réalisation de la base de données. Celles à 1 : 20 000 ou 1 : 25 000 contiennent le plus d'informations relatives à l'occupation du sol et les cartes à 1 : 50 000 sont largement utilisées pour le contrôle sur le terrain,

²Institut Géographique National : l'IGN édite des cartes topographiques par l'utilisation de moyens numériques.

- d'imagerie aérienne à des résolutions variées (noir et blanc ou infra-rouge fausses couleurs),
- de statistiques d'occupation des sols,
- d'expertises et de données ancillaires : vérité terrain, cartes des types de peuplements forestiers, cartes d'évaluation biologique, géologiques et agronomiques, relevés de température, esquisses pédologiques, cartes de l'Atlas...

Il en résulte des cartes thématiques géoréférencées et indexées reproduisant l'occupation des sols. La figure 4.1 montre un exemple de carte CLC de la région du Loiret (France), et la légende associée. A l'IFEN³ comme dans les autres institutions européennes, 5 étapes sont nécessaires pour la constitution de ces cartes :

1. L'image satellitaire est reproduite à partir d'une composition colorée dite "fausses-couleurs" sur un tirage photographique à l'échelle de 1/100 000,
2. La photo-interprétation : l'image satellitaire est interprétée visuellement en s'aidant de données exogènes (photographies aériennes, cartes topographiques IGN). Les contours des zones homogènes d'occupation du sol sont reportés sur un calque.
3. La numérisation : le calque final ou "document d'auteur" dont l'emprise correspond à une coupure 1 : 100 000 IGN, est ensuite mis au propre, contrôlé et numérisé. Chaque contour de zone est géoréférencé. L'opérateur de saisie lui affecte ensuite le code d'occupation du sol figurant sur le calque.
4. L'assemblage : les feuilles sous leur forme numérique sont alors assemblées entre elles : les bords de feuilles disparaissent.
5. La finalisation : l'équipe de production procède à une dernière vérification en superposant le document d'auteur à la restitution colorée correspondante de la base de données. Cette dernière est alors disponible sous forme numérique pour sa diffusion et son édition.

A.2 Nomenclature standard complète

Le programme CORINE Land Cover repose sur une nomenclature standard hiérarchisée à 3 niveaux et 44 postes répartis selon 5 grands types d'occupation du territoire :

1. Territoires artificialisés
 - 1.1 Zones urbanisées
 - 1.1.1 Tissu urbain continu
 - 1.1.2 Tissu urbain discontinu
 - 1.2 Zones industrielles ou commerciales et réseaux de communication
 - 1.2.1 Zones industrielles et commerciales
 - 1.2.2 Réseaux routiers et ferroviaires et espaces associés
 - 1.2.3 Zones portuaires
 - 1.2.4 Aéroports
 - 1.3 Mines, décharges et chantiers
 - 1.3.1 Extraction de matériaux
 - 1.3.2 Décharges
 - 1.3.3 Chantiers

³L'IFEN : Institut Français de l'Environnement, est chargé d'assurer la production, la maintenance et la diffusion des cartes CORINE Land Cover en France.

- 1.4 Espaces verts artificialisés, non agricoles
 - 1.4.1 Espaces verts urbains
 - 1.4.2 Equipements sportifs et de loisirs
- 2. Territoires agricoles
 - 2.1 Terres arables
 - 2.1.1 Terres arables hors périmètres d'irrigation
 - 2.1.2 Périmètres irrigués en permanence
 - 2.1.3 Rizières
 - 2.2 Cultures permanentes
 - 2.2.1 Vignobles
 - 2.2.2 Vergers et petits fruits
 - 2.2.3 Oliveraies
 - 2.3 Prairies
 - 2.3.1 Prairies
 - 2.4 Zones agricoles hétérogènes
 - 2.4.1 Cultures annuelles associées aux cultures permanentes
 - 2.4.2 Systèmes culturaux et parcellaires complexes
 - 2.4.3 Surfaces essentiellement agricoles, interrompues par des espaces naturels importants
 - 2.4.4 Territoires agro-forestiers
- 3. Forêts et milieux semi-naturels
 - 3.1 Forêts
 - 3.1.1 Forêts de feuillus
 - 3.1.2 Forêts de conifères
 - 3.1.3 Forêts mélangées
 - 3.2 Milieux à végétation arbustive et / ou herbacée
 - 3.2.1 Pelouses et pâturages naturels
 - 3.2.2 Landes et broussailles
 - 3.2.3 Végétation sclérophylle
 - 3.2.4 Forêt et végétation arbustive en mutation
 - 3.3 Espaces ouverts, sans ou avec peu de végétation
 - 3.3.1 Plages, dunes et sable
 - 3.3.2 Roches nues
 - 3.3.3 Végétation clairsemée
 - 3.3.4 Zones incendiées
 - 3.3.5 Glaciers et neiges éternnelles
- 4. Zones humides
 - 4.1 Zones humides intérieures
 - 4.1.1 Marais intérieurs
 - 4.1.2 Tourbières
 - 4.2 Zones humides maritimes
 - 4.2.1 Marais maritimes
 - 4.2.2 Marais salants
 - 4.2.3 Zones intertidales
- 5. Surfaces en eau
 - 5.1 Eaux continentales
 - 5.1.1 Cours et voies d'eau

- 5.1.2 Plans d'eau
- 5.2 Eaux maritimes
- 5.2.1 Lagunes littorales
- 5.2.2 Estuaires
- 5.2.3 Mers et océans

A.3 Matrices de confusion

A.4 Classification : données issues de scènes différentes

La région utilisée pour l'apprentissage est la *région 1*, et la zone de tests choisie est issue de la scène géoréférencée 41-253 du Loiret, prise un jour plus tard (voir section 4.1.2). Elle est délimitée en vert sur la figure A.1.

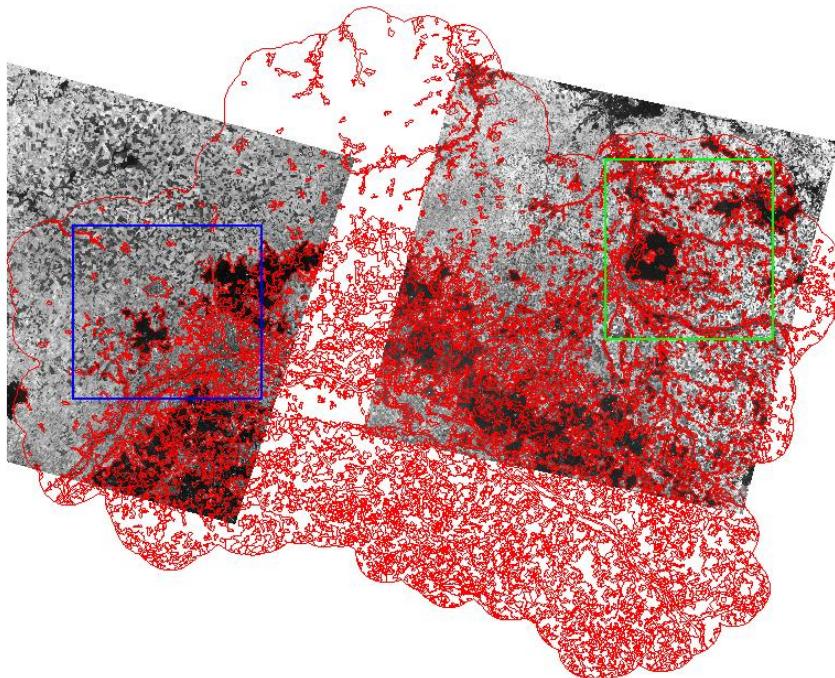


FIG. A.1 – Couche CLC vectorielle de la région du Loiret superposée à deux scènes SPOT (39-253 à gauche et 41-253 à droite). Le rectangle bleu représente la zone utilisée pour l'apprentissage. La région de tests est délimitée en vert.

Cette région de tests (nommée *région 2*) a une taille de 1340×1200 pixels (c'est-à-dire 1608000 pixels) et contient 17 classes CLC réparties comme indiqué dans le tableau A.6. Rappelons que la taille de chaque pixels est de $20m \times 20m$.

Notons l'absence des classes 221 (*Vignobles*), 122 (*Réseaux routiers et ferroviaires*), 124 (*aéroports*), 331 (*Plages, dunes et sable*) et 511 (*Cours et voies d'eau*). Comme dans la première région, la classe *terres arables* est la plus peuplée, suivie des *forêts de feuillus* et des *surfaces essentiellement agricoles interrompues par des espaces naturels importants*. La plus petite classe (*espaces verts urbains*) n'a que 243 échantillons.

TAB. A.1 – Idem avec le MAP+voisinage et les primitives {EP2}.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	43	8	2						15		5				1	2					46	
112	2	50	1						11		4				3	3	10				1	1
121	6	37	15					2	11		1				1							
122	36	41	2	5					15		23				2						2	
124	3	8	3		1				25		3				16		13				4	
131	12	2				31			2												2	
141	11								14		5				4						4	
142	4	6				1									13	61	7				5	
211	1	5				1			67	1	4				4					9	4	
221	1								4		15				1	2					21	
222		2							14		7				6	54						
231		1				1				31		39				13	49				11	
242	9					1			29		13				2	4	10	11		7		
243	5					1			16		6				1	7	41	26		14		
311		1							2		1				82	11	2			17		
312											2				39	5						
313											1				1	73	8					
324											3				2	63	12					
331	5	5				1					8				4	26	1	32		4		
411										18		1			36	8		1			36	
511	1	3								1		3			1	9					81	
512										7		2			2	3					9	

TAB. A.2 – Idem avec le MAP+voisinage et les primitives {EP3}.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	91								1						2					2		
112	4								38						2	30				10		
121	6								39	9					11	2	16	1	5		1	
122	32								35	3					3	6	2			2		
124	3								10	34					7	13	31			31		
131	17								6	15					33	4	2			2		
141	3								6	2					2	11	72	1	4		1	
142	5								6	14					11	23	30	1	5	4	1	
211	1								41	2	11				24	1	3		11		5	
221									1	20					4	68	1	6				
222	1								5	1	17				2	13	50	1	9		1	
231									21	47	1				8	14	2	8				
242	1								17	28					6	3	32	11				
243	1								10	13					6	7	46	14	1	1	1	
311									1	1	1				76	11	8	1				
312										2					14	5	78					
313										3	1				1	41	9	44				
324										4	8				34	50	9	38				
331	14															35	35	4				
411									10	1	5				33	10	1	2		38		
511	23								1	1	5				1	12	1			55		
512	82								2	2	5				4	3	3			3		

TAB. A.3 – *Région 1* : matrice de confusion obtenue en appliquant le MAP sur les 3 groupes {112, 121, 142}, {211, 231, 242, 243}, et {222, 311, 312, 313, 324}, et les classes restantes. Les caractéristiques utilisées sont {EP2} et qmf328; $\bar{T}_M=0.7653$

clc	111	112+121 +142 = 121	122	124	131	141	211+231 +242+243 = 211	221	222+311+312 +313+324 = 324	331	411	511	512
111	96	3											1
112+121 +142 = 121	1	53	8	10	8	15			2				2
122		11	87	1			1						
124		4	83				11						1
131		1	2	86			7						3
141		10	1	82					5				2
211+231+242 +243 = 211		6	9	1	79				2				2
221							99	1					
222+311+312 +313+324 = 324		2	3	2		5			82				5
331		3		10	1					73	10	4	
411		3	4		4			7			2	93	
511				3	2	2					2	84	2
512											2	91	

A. CORINE LAND COVER

TAB. A.4 – Région 1- deuxième série de tests : matrice de confusion avec le MAP + voisinage et ($\{\text{EP2}\}^+$ qmf328) pour les 22 classes; l'apprentissage a été effectué sur un cinquième des échantillons, et les tests sur les échantillons restants; $\bar{T}_M=0.6378$.

clc	111	112	121	122	124	131	141	142	211	221	222	231	242	243	311	312	313	324	331	411	511	512
111	96	2																		1	1	1
112	1	34	30	6	7	7	1	8							3					1	1	1
121	1	9	65	4	8	2	1	5							1					1	1	1
122		4	12	81		1														1	1	1
124	1	6		80		1	1	9							1					2	3	1
131		1		86		1	7													1	1	1
141	7	8			77		1	7												1	1	3
142		9		10		43	5	1							3					17	7	1
211		2	5	7		1	80		1						1	1	1	1	1	1	1	1
221							99								1	1	1	1	1	1	1	1
222		4	4		20			5	3	13	9	6			4	1	27			3		
231		9	8		16			4	40	4	4	1			3	3	1	7				
242		13	17	1	18			8	3	15	1	2	15		1		4			1	1	
243		8	18		12	1	6	3	18	4	4	6	2	4	1	14				1	3	
311		1		2		1	1	4		1		1			37	2	1	45		3	1	
312			1		1							1			4	42	50					
313			5		4		3		2			1			33	19	5	27				
324		1	1			10		15							2	95				62	8	1
331		1	2																			
411	2	1	3						7										3	1	90	1
511																				79	1	
512																					90	

TAB. A.5 – Région 1-deuxième série de tests : matrice de confusion obtenue en appliquant le MAP sur les 3 groupes 112, 124 et 142, 211, 231, 242 et 243, et 222, 311, 312, 313 et 324 et les classes restantes. Les caractéristiques utilisées sont {EP2} et qmf328; $T_M=0.7427$

clc	111	112+121 +142 = 121	122	124	131	141	211+231 +242+243 = 211	221	222+311+312 +313+324 = 324	331	411	511	512
111	96	1											3
112+121 +142 = 121	1	57	7	8	9	13		1					2
122		13	84		2			1					
124		7	81				10						1
131		1	1	84			8						4
141		12	1	86									1
211+231+242 +243 = 211		8	1	8	2	77			2				1
221			1		1			98					
222+311+312 +313+324 = 324		11		3	3	8			71				1
331		1		9	1						76		4
411		2	5	3	1	9			4		2	88	9
511				3	2	2					2	85	2
512											2	91	

TAB. A.6 – Les 17 classes CLC présentes dans la *région 2*, ainsi que le nombre de pixels pour chaque classe.

CLC	Nomenclature	Nombre de pixels
111	Tissu urbain continu	977
112	Tissu urbain discontinu	74423
121	Zones industrielles et commerciales	5992
131	Extraction de matériaux	7073
141	Espaces verts urbains	243
142	Equipements sportifs et de loisirs	4274
211	Terres arables hors périmètres d'irrigation	956369
222	Vergers et petits fruits	1147
231	Prairies	43039
242	Systèmes culturaux et parcellaires complexes	24429
243	Surfaces essentiellement agricoles interrompues par des espaces naturels importants	100639
311	Forêts de feuillus	348042
312	Forêts de conifères	19354
313	Forêts mélangées	4342
324	Forêt et végétation arbustive en mutation	1882
411	Marais intérieurs	1204
512	Plans d'eau	9571

Dans la suite, sont présentés les résultats des mêmes tests que précédemment, mais appliqués à la *région 2*, l'apprentissage étant effectué sur la *région 1*.

Le tableau A.7 présente les résultats de la classification par le Maximum a Posteriori, suivi de la prise en compte du voisinage, avec l'ensemble de caractéristiques {EP1} + qmf328).

Nous remarquons que, mis à part la classe 211 (*terres arables*) qui est assez bien classifiée mais avec beaucoup de fausses alarmes (classe majoritaire avec 956369 échantillons), toutes les autres classes sont peu ou pas reconnues, quelles que soient les caractéristiques utilisées. Par rapport à la *région 1*, les résultats sont assez mauvais.

Ceci était prévisible car c'est la première région qui a été utilisée pour l'apprentissage ; mais surtout, les deux régions appartiennent à deux scènes différentes (39-253 et 41-253) qui ont été acquises avec 1 jour d'intervalle, sous des conditions atmosphériques, radiométriques et géométriques différentes. Ces images ont subi différentes corrections (adaptation de dynamique), et pour pouvoir les comparer, et donc comparer les résultats de classification, il est nécessaire d'utiliser des réflectances qui tiennent compte de ces différences, au lieu des niveaux de gris.

Précédemment, pour le calcul des néocanaux NDVI, IB et ISU (section 4.2.3), nous avions utilisé les niveaux de gris car nous ne disposions pas des réflectances, ni des coefficients (gain, angle, ...) nécessaires pour le passage en réflectance. Cependant ces canaux ne permettent des classifications généralisables que s'ils sont faits sur des mesures ayant une véritable constance physique.

Par conséquent, l'utilisation des résultats d'apprentissage de la *région 1* pour effectuer des tests sur la *région 2* est inappropriée, ce qui explique certainement la médiocrité de nos résultats.

TAB. A.7 – *Région 2* : matrice de confusion avec le MAP + voisinage et l’ensemble de primitives ($\{\text{EP1}\}$ + qmf328) pour les 17 classes ; $\hat{T}_M=0.5949$.

Annexe B

Descripteurs statistiques de Haralick

Nous définissons ici les caractéristiques de Haralick, présentées dans la section 3.2.2, pour une image de taille $N_x \times N_y$. Ces descripteurs sont issus de matrices de co-occurrence $P(i, j, \delta, \theta)$ (équation 3.5), dont chaque élément représente une estimation de la probabilité qu'une paire de niveaux de gris i et j soit trouvée en des pixels distants de δ pixels, et ayant un angle θ par rapport à l'axe horizontal. Soit $p(i, j)$, un élément normalisé d'une matrice de co-occurrence à δ et θ fixés : $p(i, j) = P(i, j) / \sum_{i,j} P(i, j)$, on note :

$$\begin{aligned} p_y(j) &= \sum_{i=1}^{N_g} p(i, j) \\ p_{x+y}(k) &= \sum_{i=1}^{N_g} \sum_{j=1, i+j=k}^{N_g} p(i, j), k = 2, 3, \dots, 2N_g \\ p_{x-y}(k) &= \sum_{i=1}^{N_g} \sum_{j=1, |i-j|=k}^{N_g} p(i, j), k = 0, 1, \dots, N_g - 1 \end{aligned}$$

où N_g est le nombre de niveaux de gris dans l'image.

Les caractéristiques de Haralick sont définies par :

1) Moment angulaire du second ordre

$$f_1 = \sum_i \sum_j \{p(i, j)\}^2$$

2) Contraste

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \{p_{x-y}(n)\}$$

3) Corrélation

$$f_3 = \frac{\sum_i \sum_j (ij)p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

où μ_x , μ_y , σ_x and σ_y sont les moyennes et écarts-types de p_x et p_y .

4) Variance

$$f_4 = \sum_i \sum_j (i - \mu)^2 p(i, j)$$

5) Moment de différence inverse

$$f_5 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$$

6) Moyenne de somme

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$$

7) Variance de somme

$$f_7 = \sum_{i=2}^{2N_g} (i - f_6)^2 p_{x+y}(i)$$

8) Entropie de somme

$$f_8 = - \sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\}$$

9) Entropie

$$f_9 = - \sum_i \sum_j p(i, j) \log(p(i, j))$$

10) Variance de différence

$$f_{10} = \text{variance de } p_{x-y}$$

11) Entropie de différence

$$f_{11} = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$$

12,13) Mesures d'information de corrélation

$$f_{12} = \frac{H(X, Y) - H(X, Y)_1}{\max\{H(X), H(Y)\}} \quad (\text{B.1})$$

$$f_{13} = (1 - \exp[-2(H(X, Y)_2 - H(X, Y))])^{1/2} \quad (\text{B.2})$$

où $H(X, Y)$, $H(X)$ et $H(Y)$ sont les entropies, et

$$H(X, Y)_1 = - \sum_i \sum_j p(i, j) \log\{p_x(i)p_y(j)\} \quad (\text{B.3})$$

$$H(X, Y)_2 = - \sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\} \quad (\text{B.4})$$

Bibliographie

- Abdellaoui, A. & Rougab, A. (1997). Caractérisation de la réponse du bâti : application au complexe urbain de Blida (Algérie), in A. UREF (ed.), *Télédétection des milieux urbains et périurbains*, pp. 47–64.
- Adams, R. & Bishof, L. (1994). Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** : 641–647.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *Second International Symposium on Information Theory*, Budapest, pp. 267–281.
- Aksoy, S., Koperski, K., Tusk, C., Marchisio, G. & Tilton, J. (2005). Learning bayesian classifiers for scene classification with a visual grammar, *IEEE Transactions on Geoscience and Remote Sensing* **43**(3) : 581 – 589.
- Aldous, D. J. (1985). Exchangeability and related topics, in Springer (ed.), *Ecole d'été de probabilités de Saint-Flour, XIII–1983*, pp. 1 – 198.
- Alpaydin, E. (2004). *Introduction to machine learning*, MIT Press.
- Bannari, A., Morin, D. & He, D.-C. (1997). Caractérisation de l'environnement urbain à l'aide des indices de végétation dérivés des données de hautes résolutions spatiale et spectrale, in A. UREF (ed.), *Télédétection des milieux urbains et périurbains*, pp. 47–64.
- Baraldi, A. & Parmiggiani, F. (1990). Urban area classification by multispectral spot images, *Geoscience and Remote Sensing, IEEE Transactions on* **28**(4) : 674–680.
- Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. & Jordan, M. (2003). Matching words and pictures, *Journal of Machine Learning Research* **3** : 1107 – 1135.
- Bellman, R. (1961). *Adaptive Control Processes*, Princeton University Press, Princeton, NJ.
- Berge, C. (1973). *Introduction à la théorie des hypergraphes*, Les presses de l'université de Montréal.
- Beucher, S. & Meyer, F. (1993). The morphological approach to segmentation : the watershed transformation, in E. Dougherty (ed.), *Mathematical Morphology in Image Processing*, pp. 433–481.
- Bischof, H., Schneider, W. & Pinz, A. J. (1992). Multispectral classification of landsat images using neural networks, *Geoscience and Remote Sensing, IEEE Transactions on* **30**(3) : 482–490.

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer.
- Blanz, W. E. & Gish, S. L. (1990). A connectionist classifier architecture, applied to image segmentation, *In Proc. onf the 10th International Conference on pattern Recognition*, MIT Press, pp. 272–277.
- Blei, D., Griffiths, M. J. T. & Tenenbaum, J. (2003a). Hierarchical topic models and the nested chinese restaurant process, *In Advances in Neural Information Processing Systems 16 (NIPS)*, MIT Press.
- Blei, D. M. & Jordan, M. I. (2003). Modeling annotated data, *In Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '03*, ACM Press, New York, NY, USA, pp. 127–134.
- Blei, D. M. & Lafferty, J. D. (2006). Correlated topic model, *In Advances in Neural Information Processing Systems 18*.
- Blei, D., Ng, A. & Jordan, M. (2003b). Latent dirichlet allocation, *Journal of Machine Learning Research* **3** : 993 – 1022.
- Bloch, I. (2005). Fuzzy Spatial Relationships for Image Processing and Interpretation : A Review, *Image and Vision Computing* **23**(2) : 89 – 110.
- Bloch, I., Maître, H. & Anvari, M. (1997). Fuzzy adjacency between image objects, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **5**(6) : 615 – 653.
- Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning, *Artificial Intelligence* **97**(1) : 245–271.
- Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. (1987). Occam's razor, *Inf. Process. Lett.* **24**(6) : 377–380.
- Bogdan, M. (1999). Data-driven smooth tests for bivariate normality, *Journal of Multivariate Analysis* **68**(1) : 26 – 53.
- Bolon, P., Chassery, J., Cocquerez, J., Demigny, D., Graffigne, C., Montanvert, A., Philipp, S., Zeboudj, R. & Zerubia, J. (1995). *Analyse d'images : Filtrage Et Segmentation*, Masson.
- Bosch, A., Zisserman, A. & Munoz, X. (2006). Scene classification via plsa, *Proceedings of the European Conference on Computer Vision*, Vol. 4, Graz, Autriche, pp. 517–530.
- Bossard, M., Feranec, J. & Otahel, J. (2000). Corine land cover technical guide - addendum 2000, *European Environment Agency*.
- Bouillon, P. & Vandooren, F. (1998). *Traitemet automatique des langues naturelles*, De Boeck Université.
- Bozdogan, H. (1987). Model selection and akaike's information criterion (aic) : the general theory and its analytical extensions, *Psychometrika* **52** : 345–370.

- Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*, Chapman and Hall (Wadsworth, Inc), New York.
- Bretto, A. & Ubéda, S. (1996). Hypergraph model of digital topology for grey level images, *DCGI '96 : Proceedings of the 6th International Workshop on Discrete Geometry for Computer Imagery*, Springer-Verlag, London, UK, pp. 217–226.
- Burnham, K. P. & Anderson, D. R. (2002). *model selection and multi-model inference*, Springer-Verlag, New York, LLC.
- Buttner, G., Feranec, J., Jaffrain, G., Mari, L., Maucha, G. & Soukup, T. (2004). The CORINE Land Cover 2000 project, *EARSeL Proceedings*, Vol. 3, Istanbul, Turkey, pp. 331–346.
- Camastra, F. & Vinciarelli, A. (2008). *Machine Learning for Audio, Image and Video Analysis : Theory and Applications*, Springer.
- Campedel, M. & Moulines, E. (2005). Classification et sélection de caractéristiques de textures, *Revue d'intelligence artificielle* **19**(4 - 5) : 633–659.
- Campedel, M., Kyrgyzov, I. & Maître, H. (2008). Consensual clustering for unsupervised feature selection. application to spot5 satellite image indexing, *JMLR : Workshop and Conference Proceedings*, number 4, Antwerp, Belgium, pp. 48–59.
- Campedel, M., Luo, B., Kyrgyzov, I., Roux, M. & Maître, H. (December 2004). Indexation des images satellitaires : Détection et évaluation des caractéristiques de classification, *Technical Report 2004D008*, ENST Paris.
- Canters, F. (1997). Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification, *Photogrammetric engineering and remote sensing* **63**(4) : 403–414.
- Carlotto, M. J. (1998). Spectral shape classification of landsat thematic mapper imagery, *Photogrammetric engineering and remote sensing* **64**(9) : 905–913.
- Casals-Carrasco, P., Kubo, S. & Babu Madhavan, B. (2000). Application of spectral mixture analysis for terrain evaluation studies, *International Journal of Remote Sensing* **21**(16) : 3039 – 3055.
- Celeux, G. & Diebolt, D. (1985). The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly* **2**(1) : 73–82.
- Cetin, H., Warner, T. A. & Levandowski, D. W. (1993). Data classification, visualization, and enhancement using n-dimensional probability density functions (npdf) : Aviris, tims, tm, and geophysical applications, *Photogrammetric engineering and remote sensing* **59**(12) : 1755–1764.
- Chang, C. & Lin, C. (2004). Libsvm 2.5. Available from : <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Charou, E., Petridis, S., Stefouli, M., Mavranta, O. & Perantonis, S. (2005). Innovative feature selection used in multispectral imagery classification for water quality monitoring, *In XXth ISPRS Congress*, number 1398, Istanbul, Turkey, pp. 1354–1358.

- Clementini, E. & Felice, P. D. (1997). Approximate Topological Relations, *International Journal of Approximate Reasoning* **16** : 173 – 204.
- Coggins, J. M. (1982). *A framework for texture analysis based on spatial filtering*, PhD thesis, Computer Science Department, Michigan State University, East Lansing, Michigan.
- Comaniciu, D. & Meer, P. (2002). Mean shift : A robust approach toward feature space analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5) : 603–619.
- Cortes, C. & Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3) : 273–297.
- Cross, G. C. & Jain, A. K. (1983). Markov random fields texture models, *IEEE transactions on Pattern Analysis and Machine Intelligence* **5**(1) : 25–39.
- Datcu, M., Daschiel, H., Pelizzari, A., Quartulli, M., Galoppo, A., Colapicchioni, A., Pastori, M., Seidel, K., Marchetti, P. & D'Elia, S. (2003). Information mining in remote sensing image archives : system concepts, *Geoscience and Remote Sensing, IEEE Transactions on* **41**(12) : 2923–2936.
- Dean, A. M. & Smith, G. M. (2003). An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities, *International Journal of Remote Sensing* **24**(14) : 2905 – 2920.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science* **41** : 391–407.
- Dell'Acqua, F., Gamba, P., Ferrari, A., Palmason, J., Benediktsson, J. & Arnason, K. (2004). Exploiting spectral and spatial information in hyperspectral urban data with high resolution, *Geoscience and Remote Sensing Letters, IEEE* **1**(4) : 322–326.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the royal statistical society. Series B (methodological)* **39**(1) : 1 – 38.
- Deng, H. & Clausi, D. A. (2004). Unsupervised image segmentation using a simple mrf model with a new implementation scheme, *Pattern Recognition* **37**(1) : 2323–2335.
- Derin, H. & Elliot, H. (1987). Modelling and segmentation of noisy and textured images using gibbs random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**(1) : 39–55.
- Dubuisson, B. (1990). *Diagnostic et reconnaissance des formes*, Hermès, Paris.
- Duda, R. O., Hart, P. E. & Stork, D. G. (2000). *Pattern classification (2nd edition)*, Wiley - Interscience.
- Dunn, D. & Higgins, W. E. (1995). Optimal gabor filters for texture segmentation, *Image Processing, IEEE Transactions on* **4**(7) : 947–964.
- Dunn, D., Higgins, W. E. & Wakeley, J. (1994). Texture segmentation using 2-d gabor elementary functions, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **16**(2) : 130–149.

- Duygulu, P., Barnard, K., de Freitas, J. & Forsyth, D. (2002). Object recognition as machine translation : learning a lexicon for a fixed image vocabulary, *Seventh European Conference on Computer Vision IV* : 97 – 112.
- Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery : an overview, *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 1–34.
- Fei-Fei, L. & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, Washington, DC, USA, pp. 524–531.
- Forsyth, D. A. & Ponce, J. (2003). *Computer vision : a modern approach*, Prentice Hall.
- Fournier, J.-C. (2007). *Graphes et applications - Volume 1*, Hermès - Lavoisier.
- Frawley, W. J., Shapiro, P. G. & Matheus, C. J. (1992). Knowledge discovery in databases - an overview, *AI Magazine* **13** : 57–70.
- Freeman, J. (1975). The modelling of spatial relations, *Computer Graphics and Image Processing* **4**(2) : 156 – 171.
- Freksa, C. & Zimmerman, K. (1992). On the utilization of spatial structures for cognitively plausible and efficient reasoning, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Chicago, pp. 261–266.
- Frizelle, B. G. & Moody, A. (2001). Mapping continuous distributions of land cover : A comparison of maximum-likelihood estimation and artificial neural networks, *Photogrammetric engineering and remote sensing* **67**(6) : 693–705.
- Fua, P. & Hanson, A. J. (1987). Resegmentation using generic shape : Locating general cultural objects, *Pattern Recogn. Lett.* **5**(3) : 243–252.
- Giraudon, G., Garnesson, P. & ONTESINOS, P. M. (1992). MESSIE : un système multi-spécialistes en vision. Application à l’interprétation en imagerie aérienne, *Traitemet du signal* **9**(5) : 403–419.
- Gong, P. & Howarth, P. J. (1990). The use of structural information for improving land-cover classification accuracies at the rural-urban fringe, *Photogrammetric engineering and remote sensing* **56**(1) : 67–73.
- Gong, P., Marceau, D. & Howarth, P. J. (1992). A comparison of spatial feature extraction algorithms for land use classification with SPOT HRV data, *Remote Sensing of Environment* **40** : 137–151.
- Gorte, B. & Stein, A. (1998). Bayesian classification and class area estimation of satellite images using stratification, *Geoscience and Remote Sensing, IEEE Transactions on* **36**(3) : 803–812.
- Govaert, G. & Nadif, M. (2005). An em algorithm for the block mixture model, *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(4) : 643–647.

- Greenhill, D., Ripke, L., Hitchman, A., Jones, G. & Wilkinson, G. (2003). Characterization of suburban areas for land use planning using landscape ecological indicators derived from ikonos-2 multispectral imagery, *Geoscience and Remote Sensing, IEEE Transactions on* **41**(9) : 2015–2021.
- Gregorio, A. D. & Jansen, L. J. M. (1998). Land Cover Classification System (LCCS) : Classification concepts and user manual, *Technical report*, Food and Agriculture Organisation of the United Nations, Rome.
- Griffiths, T. L. & Steyvers, M. (2002). A probabilistic approach to semantic representation, *In Proc. of the 24th Annual Conference of the Cognitive Science Society*, pp. 381–386.
- Grigorescu, S., Petkov, N. & Kruizinga, P. (2000). A comparative study of filter based texture operators using mahalanobis distance, *Pattern Recognition, 2000. Proceedings. 15th International Conference on* **3** : 885–888.
- Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3** : 1157–1182.
- Hansen, M., Dubayah, R. & Defries, R. (1996). Classification trees : an alternative to traditional land cover classifiers, *International journal of remote sensing* **17**(5) : 1075–1081.
- Haralick, R. M., Shanmugan, K. & Dinstein, I. (1973). Textural features for image classification, *IEEE transactions on Systems, Man and Cybernetics* **3**(6) : 610–621.
- Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning : data mining, inference and prediction*, Springer, New York.
- Hastings, D. A. & Tateishi, R. (1998). Land Cover Classification : some new techniques, new source data, *Int. Society on Photogrammetry and Remote sensing* **32**(4) : 226–229.
- He, H. & Collet, C. (1999). Combining spectral and textural features for multispectral image classification with artificial neural networks, *International Archives of Photogrammetry and Remote Sensing*.
- Heinrich, G. (2008). Parameter estimation for text analysis,, *Technical report*, vsonix GmbH and University of Leipzig.
- Hofmann, T. (1999). Probabilistic latent semantic analysis, *In Proc. of Uncertainty in Artificial Intelligence, UAI'99*, pp. 289–296.
- Hsu, T.-I., Calway, A. D. & Wilson, R. (1993). Texture analysis using the multiresolution fourier transform, *In Scandinavian Conference on Image Analysis*, Tromso, Norway, pp. 823–830.
- Huang, X.-Q. & Liao, Z.-W. (2008). A labeling scheme based on markov random fields and gaussian mixture models for hyperspectral images, *Machine Learning and Cybernetics, 2008 International Conference on* **7** : 3619–3624.
- Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**(2) : 297–307.
- Ins (1995). CORINE Land Cover Belgique - Rapport final.

- Ins (2005a). CORINE Land Cover France - Fiches techniques. DMDS/UATD/FB/05-026.
- Ins (2005b). L'utilisation de CORINE Land Cover 2000. DMDS/UATD/FB/05-026.
- Jain, A. (1981). Advances in mathematical models for image processing, *Proceedings of the IEEE* **69**(5) : 502–528.
- Jarque, C. M. & Bera, A. K. (1987). A test for normality of observations and regression residuals, *International Statistical Review* **55**(2) : 163 – 172.
- Jensen, J. R. (1986). *Introductory Digital Image Processing : A Remote Sensing Perspective*, Prentice Hall, Englewood Cliffs, NJ, USA.
- Jeon, J., Lavrenko, V. & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models, *Annual ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 119 – 126.
- Jhung, Y. & Swain, P. (1994). A contextual classifier based on markov random fields and robust m-estimates, *Geoscience and Remote Sensing Symposium, 1994. IGARSS '94. Surface and Atmospheric Remote Sensing : Technologies, Data Analysis and Interpretation., International* **2** : 1169–1171.
- Jibrini, H. (2002). *Reconstruction de modèles de bâtiments à partir de données cadastrales vectorielles et d'un couple d'images aériennes à haute résolution*, PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France.
- Joachims, T. (1998). Text categorization with support vector machines : learning with many relevant features, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, Springer Verlag, Heidelberg, DE, Chemnitz, DE, pp. 137–142.
- Keramitsoglou, I., Sarimveis, H., Kiranoudis, C., Kontoes, C., Sifakis, N. & Fitoka, E. (2006). The performance of pixel window algorithms in the classification of habitats using vhsr imagery, *PandRS* **60**(4) : 225–238.
- Kindermann, R. & Snel, J. L. (1980). *Markov Random Fields and Their Applications*, American Mathematical Society.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information, *Problems of Information Transmission* **1** : 1–17.
- Kullback, S. & Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22** : 49 – 86.
- Kusaka, T., Egawa, H. & Kawata, Y. (1990). Classification of the SPOT Image using spectral and spatial features of Primitives Regions that have nearly uniform color, *IEEE Transactions on Geoscience and Remote Sensing* **28**(4) : 749–752.
- Kyrgyzov, I. O., Kyrgyzov, O. O., Maître, H. & Campedel, M. (2007). Kernel mdl to determine the number of clusters, *MLDM*, pp. 203–217.
- Landau, B. & Jackendoff, R. (1996). “what” and “where” in spatial language and spatial cognition, *Behavioral and Brain Sciences* **16** : 217 – 265.

- Larlus, D. & Jurie, F. (2008). Latent mixture vocabularies for object categorization and segmentation, *Journal of Image & Vision Computing*. to appear. Available from : <http://lear.inrialpes.fr/pubs/2008/LJ08a>.
- Lavrenko, V., Manmatha, R. & Jeon, J. (2003). A model for learning the semantic of pictures, in *Advances in Neural Information Processing systems (NIPS'03)*, MIT Press.
- Lazebnik, S., Schmid, C. & Ponce, J. (2006). Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories, *CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 2169–2178.
- Leibe, B., Mikolajczyk, K. & Schiele, B. (2006). Efficient clustering and matching for object class recognition, *Proceedings of British Machine Vision Conference*, p. II :789.
- Lersch, J. R., Iverson, A. E. & West, K. F. (1996). Segmentation of multiband imagery using minimum spanning trees, *SPIE*, Vol. 2758, pp. 10 – 18.
- Li, Y. & Bretschneider, T. (2006). Remote sensing image retrieval using a context sensitive bayesian network with relevance feedback, *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on* pp. 2461–2464.
- Loncaric, S. (1998). A survey of shape analysis techniques, *Pattern Recognition* **31**(8) : 983–1001.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* **60** : 91–110.
- Lowitz, G. E. (1983). Can a local histogram really map texture information ?, *Pattern recognition* **16**(2) : 141–147.
- Mallat, S. G. (1989). A theory for multiresolution signal decomposition : The wavelet representation, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(7) : 674–693.
- Mallat, S. G. (2003). *A wavelet tour of signal processing*, elsevier, second edn.
- Mandelbrot, B. B. (1983). *The fractal geometry of nature*, Freeman, San Francisco.
- Manthalkar, R., Biswas, P. K. & Chatterji, B. N. (2003). Rotation invariant texture classification using even symmetric gabor filters, *Pattern Recogn. Lett.* **24**(12) : 2061–2068.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika* **57**(3) : 519 – 530.
- Mark, D. M. & Egenhofer, M. J. (1994). Modeling Spatial Relations Between Lines and Regions : Combining Formal mathematical Models ans Human Subjects Testing, *Cartography and Geographic Information Systems* **21** : 195 – 212.
- Marszaek, M. & Schmid, C. (2006). Spatial weighting for bag-of-features, *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* **2** : 2118–2125.
- Matti-Gallice, C. & Collet, C. (2004). Morphologie mathématique et échelle : extraction du bâti à différentes résolutions spatiales, *Revue Internationale de Géomatique* **3/4**(5) : 441 – 463.

- Mclachlan, G. & Peel, D. (2000). *Finite mixture models*, John Wiley and Sons, New York.
- Meila, M. (2002). Comparing clusterings, *Technical Report 418*, University of Washington Statistics.
- Melgani, F. & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines, *Geoscience and Remote Sensing, IEEE Transactions on* **42**(8) : 1778–1790.
- Minka, T. & Lafferty, J. (2002). Expectation-propagation for the generative aspect model, *In Proc. of the 18th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 352–359.
- Mitchell, T. (1997). *Machine learning*, McGraw-Hill.
- Mitra, S., Pal, S. & Mitra, P. (2002). Data mining in soft computing framework : a survey, *Neural Networks, IEEE Transactions on* **13**(1) : 3–14.
- Monay, F. & Gatica-Perez, D. (2003). On image auto-annotation with latent space models, *MULTIMEDIA '03 : Proceedings of the eleventh ACM international conference on Multimedia*, ACM, New York, NY, USA, pp. 275–278.
- Mori, Y., Takahashi, H. & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words, *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*.
- Mueller, M., Segl, K. & Kaufmann, H. (2004). Edge- and region-based segmentation technique for the extraction of large, man-made objects in high resolution satellite imagery, *Pattern recognition* **37**(8) : 1619–1628.
- Myint, S. W., Lam, N. S.-N. & Tyler, J. M. (2004). Wavelets for urban spatial feature discrimination : comparisons with fractal, spatial autocorrelation, and spatial co-occurrence approaches, *Photogrammetric Engineering and Remote Sensing* **70**(7) : 803–812.
- Mäenpää, T. & Pietikäinen, M. (2004). Classification with color and texture : jointly or separately ?, *Pattern Recognition* **37**(8) : 1629–1640.
- Nadler, M. & Smith, E. P. (1993). *Pattern recognition engineering*, John Wiley and Sons, Toronto.
- Nakagawa, Y. & Rosenfeld, A. (1979). Some experiments on variable thresholding, *Pattern recognition* **11** : 191–204.
- Ortner, M., Descombes, X. & Zerubia, J. (2003). Building extraction from digital elevation models, *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '03)* **3** : 337 – 340.
- Otsu, N. (1979). A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man and Cybernetics (SMC)* **9**(1) : 62–66.
- Pal, M. & Mather, P. M. (2003). An assessment of the effectiveness of decision trees methods for land cover classification, *Remote sensing environment* **86**(4) : 554–565.

- Pal, M. & Mather, P. M. (2004). Assessment of the effectiveness of support vector machines for hyperspectral data, *Future Gener. Comput. Syst.* **20**(7) : 1215–1225.
- Pal, N. R. & Pal, S. K. (1993). A review on image segmentation techniques, *Pattern recognition* **26**(9) : 1277–1294.
- Pelletier, F. J. (1994). The principle of semantic compositionality, *Topoi* **13**(1) : 11 – 24.
- Peng, H., Long, F. & Ding, C. (2005). Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy., *IEEE Trans. Pattern Analysis and Machine Intelligence* **27**(8) : 1226–1238.
- Peng, T., Li, B. & Su, H. (2003). A remote sensing image classification method based on evidence theory and neural networks, *Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on* **1** : 240–244.
- Pesaresi, M. & Benediktsson, J. (2001). A new approach for the morphological segmentation of high-resolution satellite imagery, *Geoscience and Remote Sensing, IEEE Transactions on* **39**(2) : 309–320.
- Peura, M. & Iivarinen, J. (1997). Efficiency of simple shape descriptors, *In Aspects of Visual Form*, World Scientific, Capri, Italy, pp. 443–451.
- Poggi, G., Scarpa, G. & Zerubia, J. (2005). Supervised segmentation of remote sensing images based on a tree-structure mrf model, *IEEE Trans. Geoscience and Remote Sensing* **43**(8) : 1901–1911.
- Pony, O., Descombes, X. & Zerubia, J. (2000). Classification d'images satellitaires hyperspectrales en zone rurale et périurbaine, *Technical report*, INRIA, Sophia Antipolis.
- Quinlan, J. R. (1993). *C4.5 : programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.
- Randen, T. & Husoy, J. (1999). Filtering for texture classification : a comparative study, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **21**(4) : 291–310.
- Rissanen, J. (1978). Modelling by the shortest data description, *Automatica* **14** : 465–471.
- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length, *The Annals of Statistics* **11**(2) : 416–431.
- Romeu, J. & Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality, *Journal of Multivariate Analysis* **46**(2) : 309 – 334.
- Rondeaux, G., Steven, M. & Baret, F. (1996). Optimisation of Soil-Adjusted Vegetation Index, *Remote Sensing of Environment* **5**(2) : 95 – 107.
- Russell, S. & Norvig, P. (1995). *Artificial intelligence : a modern approach*, Prentice Hall, New York.
- Salembier, P. & Garrido, L. (2000). Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval, *IEEE Transactions on Image Processing* **9**(4) : 561–576.

- Schowengerdt, R. A. (1997). *Remote Sensing : Models and Methods for Image Processing (second edition)*, Academic Press.
- Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**(2) : 461–464.
- Serpico, S. & Roli, F. (1995). Classification of multisensor remote-sensing images by structured neural networks, *Geoscience and Remote Sensing, IEEE Transactions on* **33**(3) : 562–578.
- Shafer, G. (1976). *A mathematical theory of evidence*, Princeton University Press, Princeton, NJ, USA.
- Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(8) : 888–905.
- Shyu, C., Klaric, M., Scott, G., Barb, A., Davis, C. & Palaniappan, K. (2007). GeoIRIS : Geospatial Information Retrieval and Indexing System - Content Mining, Semantics Modeling and Complex Queries, *IEEE Transactions on Geoscience and Remote Sensing* **45**(4) : 839–852.
- Singh, M. & Singh, S. (2002). Spatial texture analysis : a comparative study, *Pattern Recognition, 2002. Proceedings. 16th International Conference on* **1** : 676–679.
- Smeulders, A., Worring, M., Santini, S., Gupta, A. & Jain, R. (2000). Content-based image retrieval at the end of the early years, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(12) : 1349–1380.
- Smith, S. P. & Jain, A. K. (1988). A test to determine the multivariate normality of a data set, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(5) : 757 – 761.
- Smits, P. C. & Dellepiane, S. G. and Schowengerdt, R. A. (1999). Quality assessment of image classification algorithms for land-cover mapping : a review and a proposal for a cost-based approach, *International journal of remote sensing* **20**(8) : 1461–1486.
- Song, C., Li, P. & Yang, F. (2006). Multivariate texture measured by local binary pattern for multispectral image classification, *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on* pp. 2145–2148.
- Srivastava, D. K. & Mudholkar, G. S. (2003). Goodness-of-fit tests for univaraite and multivariate normal models, in C. R. Rao & N.-H. R. Khattree, Elsevier (eds), *Handbook of Statistics : Statistics and Industry*, Vol. 22, pp. 869 – 906.
- Steinnocher, K., Kressler, F. & Kostl, M. (2003). Modelling population pressure in sub-urban and rural regions based on remote sensing and statistical data, *Geoscience and Remote Sensing Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International* **3** : 1969–1971.
- Stone, M. (1977). An asymptotic equivalence of choice of model by crossvalidation and akaike's criterion, *Journal of the Royal Statistical Society, Series B* **39** : 44–47.
- Sugar, C. & James, G. (2003). Finding the Number of Clusters in a Data Set : An Information Theoretic Approach, *Journal of the American Statistical Association* **98**(463) : 750–763.

- Sugiura, N. (1978). Further analysis of the data by the akaike's information criterion and the finite corrections, *Communications in Statistics : Theory and Methods* **7**(1) : 13–27.
- Swain, M. J. & Ballard, D. H. (1991). Color indexing, *International Journal of Computer Vision* **7**(1) : 11 – 32.
- Taxt, T., Flynn, P. & Jain, A. K. (1989). Segmentation of document images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(12) : 1322–1329.
- Theodoridis, S. & Koutroumbas, K. (2003). *Pattern Recognition, Second Edition*, Academic Press.
- Tréneau, A. & Colantoni, P. (2000). Regions adjacency graph applied to color image segmentation, *IEEE Transactions on image processing* **9**(4) : 735–743.
- Tuceryan, M. & Jain, A. K. (1998). Texture analysis, *The handbook of pattern recognition and computer vision (2nd edition)*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, pp. 207–248.
- Unsalan, C. & Boyer, K. L. (2004). Classifying Land Development in high-resolution satellite imagery using Hybrid Structural-Multispectral Features, *IEEE Transactions on Geoscience and Remote Sensing* **42**(12) : 2840–2850.
- Vanhamel, I., Pratikakis, I. & Sahli, H. (2003). Multiscale gradient watersheds of color images, *Image Processing, IEEE Transactions on* **12**(6) : 617–626.
- Vapnik, V. (1996). *The nature of statistical learning theory*, Springer-Verlag, New-York.
- Vetterli, M. (1986). Filter banks allowing perfect reconstruction, *Signal Process.* **10**(3) : 219–244.
- Wang, X. & Grimson, E. (2008). Spatial latent dirichlet allocation, *In Advances in Neural Information Processing Systems 20 (NIPS)*.
- Watanabe, T., Sugawara, K. & Sugihara, H. (2002). A new pattern representation scheme using data compression, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5) : 579–590.
- Weiss, S. M. & Indurkhya, N. (1998). *Predictive data mining : a practical guide*, Morgan Kaufmann, San Francisco.
- Wilkinson, G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments, *Geoscience and Remote Sensing, IEEE Transactions on* **43**(3) : 433–440.
- Wilkinson, G., Fierens, F. & Kanellopoulos, I. (1995). Integration of neural and statistical approaches in spatial data classification, *Geographical Systems* **2** : 1–20.
- Winn, J., Criminisi, A. & Minka, T. (2005). Object categorization by learned universal visual dictionary, *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* **2** : 1800–1807.
- Witten, I. H. & Frank, E. (2005). *Data mining : practical machine learning tools and techniques (second edition)*, Morgan Kaufmann.

- Wu, F. Y. (1982). The potts model, *Reviews of Modern Physics* **54**(1) : 235+.
- Yan, G., Mas, J. F., Maathuis, B. H. P., Xiangmin, Z. & Van Dijk, P. M. (2006). Comparison of pixel-based and object-oriented image classification approaches : a case study in a coal fire area, wuda, inner mongolia, china, *International Journal of Remote Sensing* **27**(18) : 4039–4055.
- Yang, C., Dong, M. & Fotouhi, F. (2005). Region based image annotation through multiple-instance learning, *MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia*, ACM, New York, NY, USA, pp. 435–438.
- Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods, *In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-99)*, pp. 42–49.
- Zadeh, L. A. (1965). Fuzzy sets, *Image and Control* **8** : 338 – 353.
- Zahzah, E.-H., Malki, J. & Mascarilla, L. (2003). Spatial relations for dynamic changes and objects retrieval, *Journal of Geographic Information and Decision Analysis* **6**(2) : 176–202.
- Zammit, O., Descombes, X. & Zerubia, J. (2007). Assessment of different classification algorithms for burnt land discrimination, *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International* pp. 3000–3003.
- Zhang, D. & Lu, G. (2004). Review of shape representation and description techniques, *Pattern Recognition* **37**(1) : 1–19.
- Zhang, J. & Foody, G. M. (1998). A fuzzy classification of sub-urban land cover from remotely sensed imagery, *International journal of remote sensing* **19**(14) : 2721–2738.
- Zhou, Q. & Robson, M. (2001). Automated rangeland vegetation cover and density estimation using ground digital images and a spectral-contextual classifier, *International Journal of Remote Sensing* **22**(17) : 3457–3470.

