Apprentissage Bayésien de Structures Spatio-Temporelles : “application à la fouille visuelle de séries temporelles d’images de satellites”

THÈSE

présentée pour obtenir

LE TITRE DE DOCTEUR DE L’ECOLE NATIONALE SUPERIEURE DE L’AERONAUTIQUE ET DE L’ESPACE

École doctorale : INFORMATIQUE ET TELECOMMUNICATIONS
Spécialité : SIGNAL, IMAGE et ACOUSTIQUE

par

Patrick HÉAS

Soutenue le 8 avril 2005 devant le jury composé de:

M. Henri MAITRE	Rapporteur
M. Ali MOHAMMAD-DJAFARI	Rapporteur
M. Alain GIROS	Examinateur
M. Manuel SAMUELIDES	Examinateur
M. Mihai DATCU	Directeur de thèse
M. Philippe MARTHON	Directeur de thèse

Résumé

Durant les dernières décennies, les satellites n'ont cessé d'acquérir des images de haute résolution de beaucoup de sites d'observation de la Terre. De nouveaux produits sont apparus avec ce processus d'acquisition intensif : les séries temporelles d'images de satellites de haute résolution. Elles représentent un important volume de données dont le riche contenu informatif est susceptible d'intéresser un large panel d'applications nouvelles.
Cette thèse présente un concept de fouille d'information qui permet l'apprentissage de structures spatio-temporelles contenues dans les séquences d'images, l'objectif étant l'interprétation et la recherche probabiliste de phénomènes dans l'espace et le temps. Les connaissances expertes d'un utilisateur conduisent le processus d'apprentissage, via la communication d'exemples et de contre exemples. Les fondements théoriques de ce concept se situent à l'interface de l'inférence bayésienne et entropique, des modèles stochastiques et de la cognition visuelle. Le concept emploie une modélisation hiérarchique bayésienne du contenu des séquences d'images, qui permet de lier les intérêt des utilisateurs aux différentes structures spatio-temporelles. La hiérarchie comprend deux principales phases d'apprentissage : l'inférence non supervisé d'un graphe de trajectoires de clusters dynamiques et, basé sur ce graphe, l'apprentissage interactif d'étiquettes sémantiques associées aux structures spatio-temporelles contenues dans la scène dynamique.
Les algorithmes et méthodes développés sont intégrés dans un système de fouille visuelle d'information. Ce système représente un outil entièrement novateur pour l'exploitation du contenu des séries temporelles d'images satellites de haute résolution. Les expériences effectuées avec une série temporelles d'images SPOT démontrent les capacités du système dans la compréhension de scènes dynamiques.

Mot-clés : Analyse spatio-temporelle, fouille d'information, modélisation bayésienne, théorie de l'information, apprentissage interactif, graphe de trajectoires de clusters, étiquetage sémantique

Abstract

During the last decades, satellites have acquired incessantly high resolution images of many Earth observation sites. New products have arisen from this intensive acquisition process : high resolution satellite image time-series. They represent a large data volume with a rich information content and may open a broad range of new applications.
This thesis presents a knowledge-driven information mining concept which enables a system to learn, interpret and retrieve spatio-temporal phenomena contained in satellite image time-series. The concept relies on the theoretical background of inference with bayesian and entropic methods, stochastic models and visual cognition. The concept is based on a bayesian hierarchical modeling of image time series information content, which links user interests to the different spatio-temporal structures. The hierarchy is composed of two inference steps : an unsupervised modeling of dynamic clusters resulting in a graph of trajectories, and an interactive learning procedure based on graphs which leads to the semantic labeling of spatio-temporal structures.
The developed algorithms and methods are integrated in a visual information mining system. This system represents a completely novel tool to exploit high resolution satellite image time series. Experiments performed on a SPOT image time-series demonstrate the concept capabilities for dynamic scene understanding.

Keywords : Spatio-temporal analysis, information mining, Bayesian modeling, information theory, interactive learning, graph of cluster trajectories, semantic labeling.

Table des matières

Table des figures

Chapitre 1

Introduction

Un des objectifs du traitement du signal est l'extraction d'information à partir de données complexes affectées par des incertitudes. L'analyse peut s'avérer être particulièrement difficile lorsque l'on considère des processus multidimensionnels variant dans l'espace et dans le temps. Ainsi, pour accomplir une extraction d'information pertinente, il convient souvent de réduire la complexité du traitement, tout en considérant les incertitudes des données. La modélisation bayésienne et la théorie de l'information apportent une réponse à cette problématique. Ces théories proposent un contexte adapté à l'extraction d'information, par le choix et l'ajustement de modèles probabilistes aux données.

Cette thèse étudie et propose des nouvelles méthodes d'apprentissage de structures spatio-temporelles, dans le contexte de la fouille d'information de séries temporelles d'images satellites. L'objectif est l'élaboration d'un concept permettant la compréhension de scènes dynamiques, l'étiquetage sémantique de structures spatio-temporelles et leurs recherches probabilistes. L'apprentissage se base sur un modèle hiérarchique bayésien, qui, à partir de données, engendre des représentations de l'information possédant des niveaux croissants d'abstraction. Ainsi, le modèle permet de lier les structures complexes et bruitées contenues dans les données aux divers intérêts des utilisateurs. L'accent est mis sur l'étude des méthodes entropiques et bayésiennes pour la sélection de modèles et l'estimation de paramètres.

Le système développé dans cette thèse est motivé par les problèmes d'indexation par le contenu, d'interprétation et d'exploitation du contenu informatif des masses grandissantes de données spatio-temporelles issues de l'observation de la Terre. Ces problèmes sont la plupart du temps spécifiques aux domaines d'application et nécessitent encore souvent des moyens humains considérables. Un système novateur de fouille visuelle d'information d'utilisation interactive, s'adaptant à la spécificité de n'importe quelle application, est donc proposé et évalué sur des séries temporelles d'images de haute résolution acquis par les satellites SPOT.

1.1 Apports de l'inférence bayésienne dans le traitement de l'information

Les techniques de modélisation bayésienne pour l'extraction d'information à partir de données incomplètes et bruitées ont été établies depuis un certain nombre d'années dans la littérature spécifique au traitement du signal et de l'image. Des méthodes ont été développées dans divers domaines, allant de l'analyse fréquentielle (*Jaynes, 1987*) à la modélisation de la conjecture d'un utilisateur pour la recherche d'images par le contenu (*Cox et al., 2000*). Des techniques ont été proposées par *Besag (1986)* pour le filtrage d'images bruitées. *Cheesman (1988)* propose l'inférence bayésienne d'un modèle de mélange de gaussiennes et utilise l'algorithme d'Espérance-Maximisation *(Dempster et al., 1977)* pour conduire le processus d'estimation. L'approche bayésienne dans l'analyse de variables multidimensionnelles est largement utilisée. Les techniques de modélisation bayésienne par les champs aléatoires de Gibbs sont introduites dans le domaine par *Geman et Geman (1984)*. Ces mêmes techniques sont employées pour la détection de structures linéaires dans des images SAR *(Tupin et al., 1998)* et pour la reconstruction 3D de scènes urbaines à partir d'images SAR très haute résolution *(Quartulli et Datcu, 2003)*. *Heckerman (1999)* propose plusieurs modèles pour l'apprentissage supervisé avec des réseaux bayésiens. Cette idée est reprise par *Schroeder et al. (2000)* et incorporée dans un modèle hiérarchique bayésien pour la recherche d'image par le contenu. Un réseau bayésien est aussi employé par *Naphade et al. (1998)*, pour modéliser la sémantique de vidéos avec une base d'apprentissage. Un nombre important de techniques, utilisées pour la modélisation hiérarchique bayésienne et pour la sélection de modèle ainsi que pour l'estimation de paramètres, sont résumées par *Datcu et al. (1998)*.

La modélisation bayésienne est souvent caractérisée par le rôle fondamental joué par la distribution a priori. L'usage d'a priori subjectifs a souvent causé beaucoup d'objections et controverses. *Jeffrey (1939)*, et plus tard *Jaynes (1968)* avec le principe du maximum d'entropie, ont posé les bases pour le développement de techniques pour la génération de distributions a priori objectives. De plus pour faire la sélection d'un modèle a priori, une inférence bayésienne à deux niveaux peut être appliquée : *MacKay (1992)* utilise la maximisation de l'évidence bayésienne pour la sélection de modèles ; par ailleurs, *Schwarz (1978)* dérive une approximation du premier ordre de l'évidence pour s'affranchir de calculs trop coûteux.

L'analogie entre la minimisation de la divergence de Kullback-Leibler, pierre angulaire des méthodes entropiques, et le formalisme bayésien est très fort. En effet, on peut interpréter la divergence comme l'extraction d'information a posteriori à partir de connaissances a priori. Cette idée est la base du principe de discrimination minimum de Kullback *(Kullback, 1966)*, du principe du maximum d'entropie et du critère d'information d'Akaike *(Akaike, 1973)*. *Mohammad-Djafari (1994)* fait une rétrospective de l'utilisation de ces techniques en traitement du signal. Concernant l'analyse multidimensionnelle, les équivalences entre l'analyse en composantes indépendantes par des méthodes entropiques et bayésiennes sont résumées par *Lee et al. (2000)*. Mais l'analogie avec l'inférence bayésienne la plus com-

plète se trouve dans le principe de la longueur minimum de description (MDL) dérivée par *Rissanen (1978, 1986)*. L'idée d'une longueur minimum de description pour la modélisation a initialement été formulée par *Solomonoff (1964)*. Indépendamment de Solomonoff et de Rissanen, *Wallace et Boulton (1968)* formulèrent l'idée très similaire d'une longueur minimum de message (MML). Ces derniers ont appliqué leur principe au problème d'inférence de modèles de mélange de distributions *(Wallace et Dowe, 1994)*. Les équivalences et approximations entre l'évidence bayésienne et la longueur minimum de description du principe MDL/MML sont démontrées lorsque les descriptions sont basées sur le codage de Shannon par *Clarke et Barron (1990)*. *Vitanyi et Li (2000)* montrent les divergences des méthodes dans le cas de descriptions basées sur la complexité algorithmique de Kolmogorov.

1.2 Le contenu des séries temporelles d'images satellites : modélisation physique versus fouille d'information

L'exploitation des séries temporelles d'images satellites constitue un enjeu majeur pour un nombre croissant de domaines d'application intéressés par la compréhension d'évolutions spatio-temporelles.

Des séries temporelles d'images de basse résolution ont été utilisées durant les dernières décennies pour surveiller et prédire des structures globales tels que l'atmosphère, l'océan ou encore la végétation. Les données satellitaires employées en entrée de modèles physiques, ont permis la résolution de problèmes inverses et la déduction de paramètres de modèles physiques, décrivant les propriétés et les comportements des structures le plus souvent au niveau pixellaire. Mais de nos jours, la disponibilité d'une nouvelle génération de séries temporelles d'images satellites de résolution spatiale beaucoup plus haute donne accès à quantité de nouvelles structures spatio-temporelles. L'application de modèles physiques sur ces dernières n'est pas aisée. En effet, la complexité des données est plus grande du fait de l'importante diversité de structures spatio-temporelles qui survient lorsque la résolution spatiale des images atteint quelques décamètres. La figure 1.1 illustre l'accroissement de la quantité d'information disponible par l'avènement de la haute résolution. De plus, la fréquence temporelle d'échantillonnage requise pour analyser ces nouvelles structures est souvent trop faible et ne permet pas une modélisation classique des séries temporelles au niveau pixellaire.

La fouille d'information constitue une approche adaptée à la discrimination de caractéristiques comportementales des différentes structures, qui s'affranchit d'une modélisation spécifique de chaque type de structure. Dans cette situation les modèles physiques sont remplacés par des modèles entropiques. Par cette approche générique, on cherche à accéder aux séries temporelles d'images par leur contenu, à reconnaître et associer des structures spatio-temporelles, c'est à dire à localiser spatialement et temporellement des phénomènes similaires. D'un point de vue plus général, on cherche à comprendre l'évolution de scènes dynamiques.

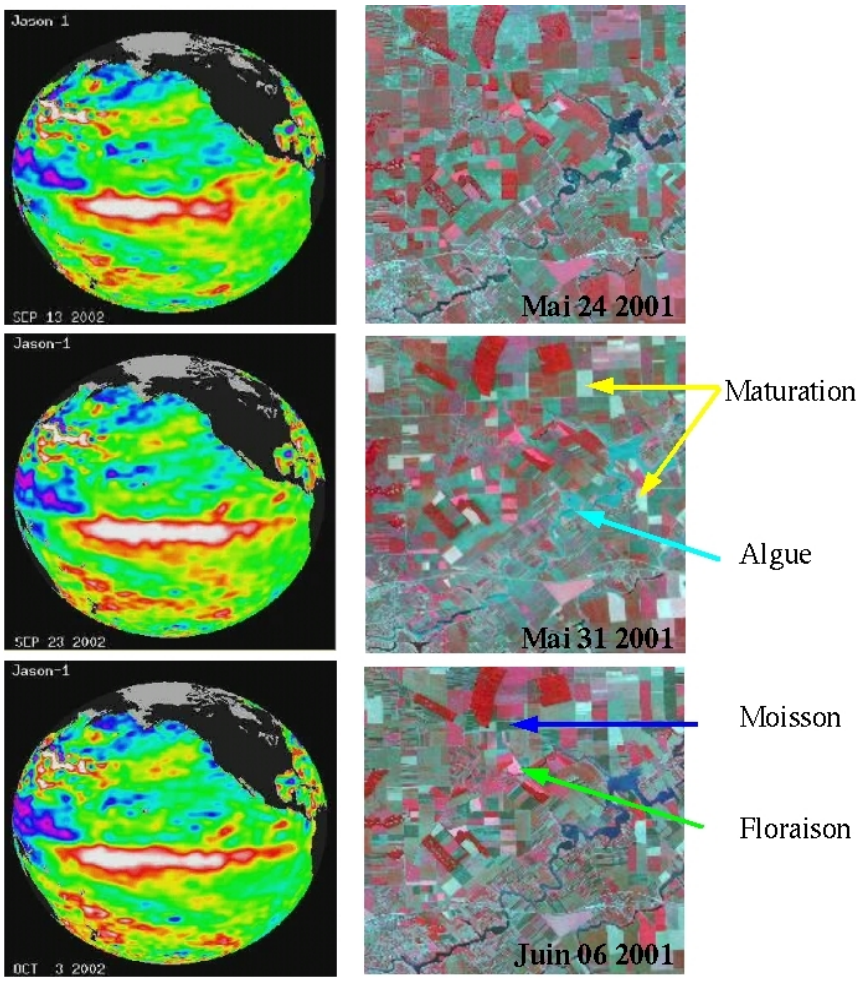


FIG. 1.1 – *Accroissement du nombre de phénomènes observables par l'avènement de la haute résolution. A droite, série temporelle d'images de basse résolution acquise par le satellite Jason 1, présentant l'évolution sur trois jours consécutifs du niveau de la surface de l'océan Pacifique ; l'analyse par des modèles physiques permet la caractérisation de phénomènes tels que "El Nino". A gauche, série temporelle d'images de haute résolution acquise par les satellites SPOT, présentant la dynamique d'une scène rurale à l'Est de Bucarest (Roumanie).*

1.3 Grandes lignes et contributions de cette thèse

Le concept développé dans cette thèse est centré sur la modélisation stochastique du contenu informatif de séries temporelles d'images satellites : un modèle hiérarchique bayésien partant des caractéristiques primitives de la série temporelle d'images et engendrant des représentations de l'information de niveau croissant d'abstraction est défini. Le but est l'apprentissage de liens probabilistes entre les différentes structures spatio-temporelles contenues dans les données et les sémantiques spécifiques aux utilisateurs (Chapitre 5). Les niveaux successifs d'apprentissage se basent sur des familles de modèles stochastiques adaptées à l'analyse de signaux multidimensionnels, sur lesquelles sont appliqués des inférences bayésiennes ou issues de méthodes entropiques *(Héas et Datcu, 2004c)*.

La modélisation hiérarchique comprend les deux principales phases d'apprentissage suivantes.

- Une représentation objective et robuste du signal spatio-temporel est engendrée par un apprentissage non-supervisé. Cette représentation est issue d'une modélisation des trajectoires de clusters dynamiques dans l'espace des caractéristiques évolutives. Les trajectoires sont codées dans des graphes, qui représentent des modèles pour les structures spatio-temporelles de la série temporelle d'images satellites (Chapitre 6). L'inférence de ces graphes s'effectue en plusieurs étapes : suite à une phase d'extraction de caractéristiques spatio-temporelles, l'information est condensée par des techniques existantes d'analyse multidimensionnelle ; le principe d'inférence MDL est adapté au cas du modèle de mélange d'un nombre indéterminé de gaussiennes pour apprendre les distributions des espaces de caractéristiques ; le processus engendre des partitions de clusters et de classes qui sont ensuite utilisées pour la modélisation de trajectoires de clusters et de classifications dynamiques *(Héas et al., 2003a)*.

- Basé sur les graphes de trajectoires, le paradigme de la fouille d'information est appliqué pour l'apprentissage interactif de la sémantique des motifs spatio-temporels de cette représentation. Une méthode est donc développée pour effectuer interactivement l'étiquetage sémantique des structures spatio-temporelles (Chapitre 7) : un modèle paramétrique évaluant la similarité entre graphes est d'abord introduit ; on dérive de celui-ci un modèle sémantique paramétrique, où une sémantique particulière correspond à un certain vecteur de paramètres ; pour estimer les paramètres du modèle et ainsi définir une sémantique, un apprentissage interactif est mis en oeuvre. Celui-ci se base sur l'application d'un modèle de Dirichlet à un réseau bayésien. Le processus d'apprentissage s'effectue via la communication d'exemples et de contre-exemples de graphes au système, qui permettent respectivement l'apprentissage d'une sémantique positive et d'une sémantique négative ; puis par la règle de Bayes, des probabilités a posteriori sont inférées pour un certain label sémantique ; le processus d'apprentissage est par ailleurs évalué après chaque nouvel exemple d'utilisateur par des mesures de pertinence *(Héas et Datcu, 2004a)*.

Les algorithmes développés sont intégrés dans un système de fouille visuelle d'information basé sur une architecture client-serveur (Chapitre 8.1). Les graphes de trajectoires et autres

modèles codant les séries temporelles d'images satellites sont sauvegardés sur le serveur et transmis sur demande à la machine client. Sur cette dernière, une interface Homme-Machine permet la navigation au travers de différentes représentations de l'information contenue dans le signal, l'interaction de l'utilisateur avec le système par la désignation d'exemples ainsi que la visualisation de l'étiquetage sémantique induit.

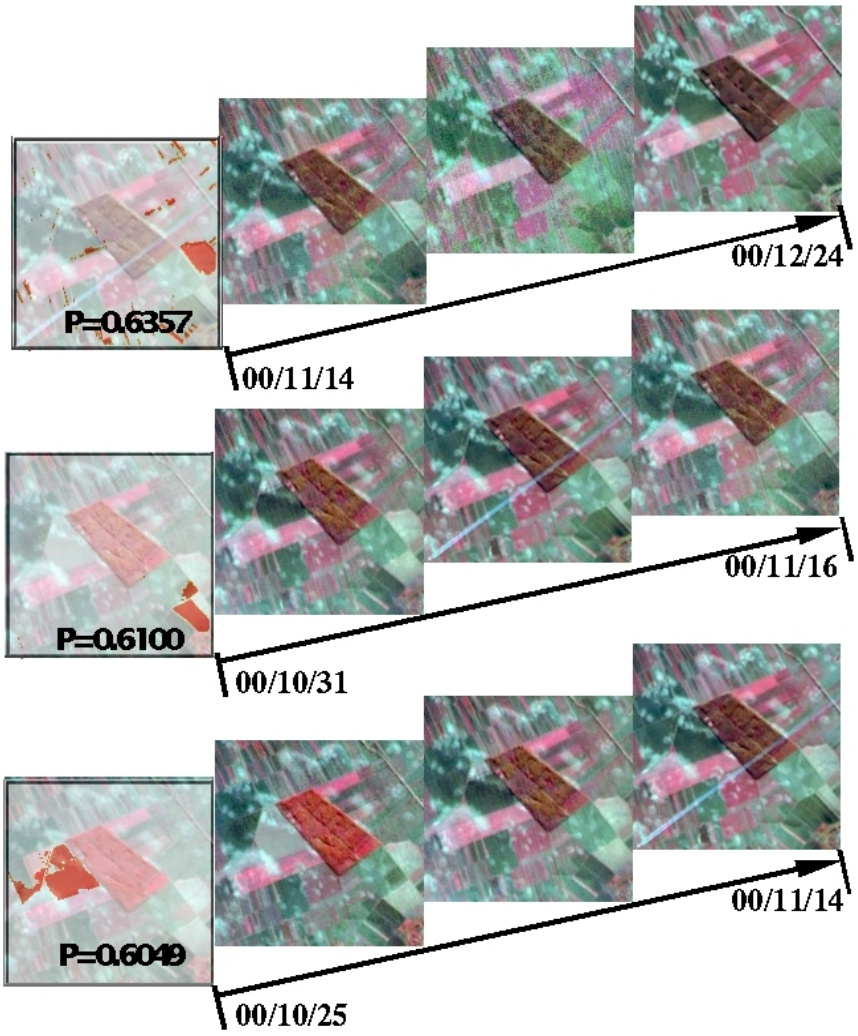


FIG. 1.2 – *Résultat d'une recherche probabiliste de phénomènes de labours ayant lieu dans une scène à l'Est de Bucarest. La recherche aboutit à la localisation dans l'espace (classes en rouge) et dans le temps (intervalles temporels annotés aux 3 séries d'images) de structures spatio-temporelles de même sémantique.*

Le système est validé sur des séries temporelles de 38 images SPOT, possédant une résolution spatiale de 20 mètres (Chapitre 8.2). Les différents éléments de la chaîne de traitement sont évalués. En particulier, les classifications obtenues sont mises en rapport avec des données "terrain" issues du projet d'Assimilation de Données par Agro-Modélistion (ADAM). Les capacités du système à apprendre interactivement des

sémantiques spécifiques d'utilisateurs sont démontrées : des phénomènes de moisson, de labour, de maturation de culture sont appris et les localisations spatio-temporelles d'événements de même sémantique sont retrouvées de manière probabiliste dans la base de données ; l'identification d'évolutions de cultures spécifiques est menée avec succès. La figure 1.2 montre un de ces résultats.

Le système développé dans cette thèse constitue un outil entièrement novateur dans le domaine de l'exploitation de séries temporelles d'images satellites de haute résolution. En effet, aucun système actuel ne considère la diversité de l'information contenue dans de telles données. La modélisation de cette grande hétérogénéité d'information est permise par une approche basée sur des fondements théoriques puissants pour le traitement de l'information. Ainsi, le concept décrit dans cette thèse permet d'étendre le paradigme de la fouille d'information aux séries temporelles d'images satellites.

Ce travail est présenté de la façon suivante : dans le chapitre 2, les séries temporelles d'images satellites, ainsi que les structures spatio-temporelles qu'elles contiennent, sont caractérisées. Le chapitre 3 présente le paradigme de la fouille d'information et dresse un état de l'art du domaine. La problématique de la thèse est explicitement formulée à la fin de ce chapitre. Dans le chapitre 4, les fondements théoriques du concept sont introduits par un exposé sur les techniques de modélisation bayésienne. Le chapitre 5 constitue le noyau de cette thèse, avec une description complète du concept d'apprentissage spatio-temporel. Les chapitres 6 et 7 détaillent plus amplement les deux grandes parties constituant la modélisation hiérarchique : l'apprentissage non-supervisé et l'apprentissage interactif. Dans le chapitre 8, l'organisation du système de fouille est décrite et une évaluation est effectuée. Finalement, la conclusion présente un résumé des travaux ainsi que les perspectives qui s'en dégagent.

Première partie

Caractérisation du problème

Chapitre 2

Les Séries Temporelles d'Images Satellites (STIS)

Dans ces travaux de thèse, les données utilisées sont les images optiques acquises par les "Satellites Pour l'Observation de la Terre" (SPOT). Elles constituent une partie de la base de données du projet "Assimilation de Données par Agro Modélisation" (ADAM). Elles relatent la dynamique d'une scène rurale de la plaine du Danube en Roumanie, à l'Est de Bucarest. L'objectif du projet ADAM, abordé dans le chapitre 3.4, est focalisé sur l'extraction d'information agronomique. Les données acquises pour ce projet sont accessibles en ligne [8].

2.1 Caractérisation des données SPOT du projet ADAM

Les satellites SPOT utilisent des capteurs appelés radiomètres qui couvrent les domaines spectraux visible et infrarouge. Les radiomètres embarqués sont calibrés pour observer 3 bandes spectrales appelées canaux : un canal dans le "vert-jaune", un canal dans le "rouge" et un dans le proche "infra-rouge" qui correspondent respectivement à des intervalles de longueurs d'onde allant de 0.50 à 0.59 μm, de 0.61 à 0.68 μm et de 0.79 à 0.89 μm.
Les satellites SPOT possèdent des capteurs à haute résolution spatiale : 1 pixel représente environ $20m$x$20m$[1] Ils défilent sur une orbite polaire héliosynchrone et sont capables d'enregistrer une image multispectrale, acquise ligne par ligne, de 3000x3000 pixels.

Pour le projet ADAM, les images optiques ont été acquises par SPOT 1, 2 et 4. L'acquisition d'images de la même scène rurale a été journalière durant une période de 10 mois, allant d'octobre 2000 à juillet 2001. Pour constituer la STIS finale, les images ont ensuite été soumises à plusieurs étapes de sélection et de prétraitement. La section suivante détaille les différents processus postérieurs à l'acquisition nécessaires à la constitution du jeu de données ADAM.

[1] En géométrie brute, un pixel projeté au sol a une taille de $20m$ en longueur et entre $20m$ et $27m$ en largeur. Ce n'est qu'après rééchantillonage, qu'on obtient des pixels de taille $20m$x$20m$.

2.2 Constitution du jeu de données ADAM

2.2.1 Sélection des images

Dans le cadre du projet "ADAM", pour mieux répondre aux besoins de l'agro-assimilation, les images acquises quotidiennement ont été sélectionnées en supprimant les images présentant une couverture neigeuse ou nuageuse sur certains sites de culture "test' de la scène. Trente huit images irrégulièrement échantillonnées dans le temps, sur une période de 286 jours résultent de cette sélection. La figure 2.1 présente l'échantillonnage irrégulier de la STIS.

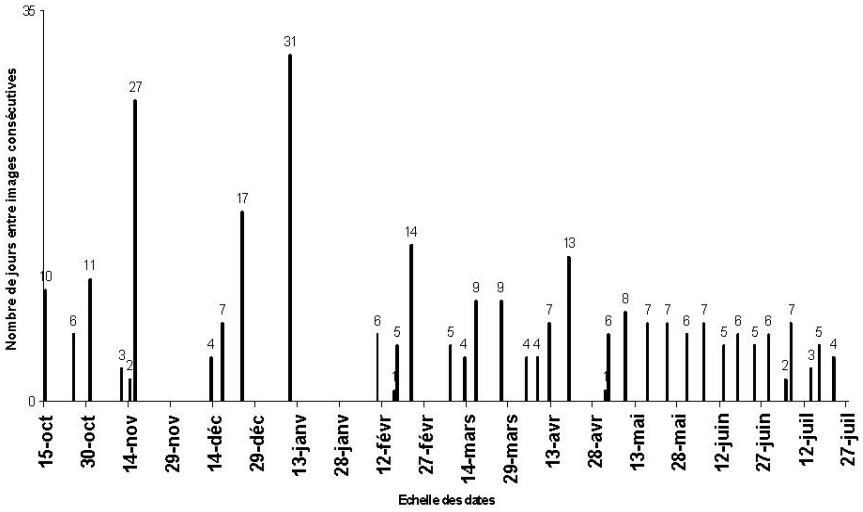


FIG. 2.1 – *Nombre de jours entre l'acquisition d'images sans couverture nuageuse ou neigeuse dans la base de données ADAM. L'axe horizontal représente le temps et supporte l'échantillonnage irrégulier des images. L'axe vertical représente la différence de temps entre les dates consécutives d'acquisition.*

2.2.2 Corrections géométriques

Afin de projeter les 38 images dans le même repère cartographique, les images acquises par les satellites sont soumises à des prétraitements géométriques.
Un modèle numérique de terrain (MNT) du site a d'abord été fabriqué à partir de deux images formant un couple stéréoscopique adéquat. Les coordonnées géographiques (latitude, longitude et altitude) d'une trentaine de points de liaisons visibles dans chacune des 38 images, mises en relation avec le MNT du site et des données d'acquisition du satellite (angles de prise de vue, altitudes du satellite, ...), ont permis, par l'intermédiaire d'une spatio-triangulation et d'interpolations, de projeter les images dans un repère géographique puis cartographique. Les images finales extraites ont une taille de 3000x2000 pixels[2] , soit

[2]En projetant les images dans un repère cartographique, la taille en pixels des images brutes (3000x3000 pixels) est convertie en une taille métrique (supérieure ou égale à 60x60km^2 en fonction des angles de prise

60km x 40km.
Par ce traitement, les images sont rendues géométriquement superposables dans la projection cartographique donnée avec une bonne précision : le diamètre maximum des cylindres, contenant pour chacun l'évolution spatiale d'un point de l'image au cours du temps, est de 1.5 pixel.

2.2.3 Corrections radiométriques

Chaque pixel de chaque image de la série temporelle a aussi été soumis à des corrections radiométriques. La radiométrie des images a dans un premier temps été corrigée par l'utilisation de modèles physiques. Dans une deuxième étape, la radiométrie des images a été ajustée par inter-calibration entre couple d'images consécutives.

Introduction d'un modèle de transfert radiatif pour le calcul des réflectances Les images acquises par les capteurs des satellites sont des mesures de luminance. Elles sont représentatives de l'énergie des radiations réfléchies par les structures, pour une gamme de longueurs d'onde et suivant un angle de visée et une surface donnée.
Les mesures de luminance sont d'abord converties en mesures de réflectance en effectuant le rapport entre l'énergie réfléchie par ces structures et l'énergie des radiations incidentes provenant du soleil, reçue par ces mêmes structures pour une longueur d'onde et une surface donnée.
Ensuite des corrections atmosphériques sont effectuées par l'intermédiaire d'un modèle de transfert radiatif prenant en compte des données exogènes telles que des mesures quantifiant la vapeur d'eau, l'ozone, et les aérosols présents aux moments de l'acquisition.
Les images résultant de ce traitement représentent des mesures physiques, pour des gammes de longueurs d'ondes données, d'énergie réfléchie par les structures en dessous de la couverture atmosphérique, qui sont indépendantes de l'angle de visée du satellite.
Mais la complexité des effets atmosphériques rend la modélisation imparfaite. En conséquence, à défaut d'avoir des mesures physiques absolues pour chaque image, des mesures relatives sont introduites dans le but d'obtenir une STIS possédant des radiométries homogènes.

Ajustement linéaire des radiométries Afin d'atténuer les transitions radiométriques brusques dans la STIS, la radiométrie des images est ajustée linéairement pour des couples d'images (I_t, I_{t+1}) successifs, pris chronologiquement. Pour ces ajustements, chaque bande spectrale est considérée indépendamment des autres. Pour chaque bande spectrale, on constitue ainsi 37 scatterogrammes. La fréquence moyenne d'échantillonnage temporel étant suffisamment haute, on fait l'hypothèse de corrélation entre une proportion suffisamment importante de pixels pour estimer une droite dominante dans chaque scatterogramme par recherche du maximum de la transformée de Radon. En formulant l'hypothèse que les pixels ayant servi à l'estimation de la droite possèdent des valeurs radiométriques inchangées entre les temps t et $t + 1$, la différence entre la pente de chaque droite estimée

de vue). Pour obtenir une STIS homogène, une portion de 3000x2000 pixels, représentant 60km x 40km, commune à toutes les images est extraite.

et la pente unité est introduite pour l'ajustement linéaire des pixels de l'image I_{t+1}.
Par ce traitement, des réflectances relatives, présentant une homogénéité temporelle, sont estimées. On peut ainsi caractériser des courbes de réflectance spectrale de différentes structures pour différentes longueurs d'onde variant avec le temps et donc, observer des évolutions de signatures spectrales particulières.

Les traitements géométriques et radiométriques détaillés plus haut permettent d'aboutir à la série temporelle présentée dans la figure 2.2 .

2.3 Structures spatio-temporelles

Les STIS sont des objets complexes contenant des structures spatio-temporelles nombreuses et variées : apparitions de nuages, de véhicules ou d'avions, croissance ou moisson de cultures parcellaires, comportement des rivières et fleuves, etc. A titre d'exemple, des processus de moissons et de labours intervenant entre le 15 octobre 2000 et le 17 février 2001 sont visibles dans la figure 2.3.
L'analyse de ces structures spatio-temporelles est utile pour l'étude et la compréhension d'évolutions complexes dans des domaines variés tels que l'agriculture, l'étude des forêts, l'océanographie, l'hydrologie, l'urbanisation, etc. Dans une zone majoritairement rurale, la compréhension de la dynamique de la scène pourra être par exemple l'identification de phénomènes temporellement décalés caractérisant un processus progressif de moissons. D'autres exemples de compréhension de scène dynamique rurale pourront être l'identification d'interaction spatiale entre des cultures et des usines polluantes ou encore l'association de variations brutales des radiométries avec des phénomènes d'occlusion par des nuages ...

L'avènement de la haute résolution donne l'accès à d'énormes quantités d'informations spatialement précises. De plus, la constitution de STIS haute résolution rend possible l'étude de la dynamique de ces structures détaillées.
Ces structures spatio-temporelles contenues dans les STIS sont caractérisées par des échelles temporelles variées comme l'illustre la figure 2.4. La scène dynamique comprend des motifs se déroulant sur une période courte comme l'occlusion d'une parcelle agricole par un nuage, le passage d'un avion ou une rupture induite par une moisson ou le labourage d'un champ. La STIS comprend aussi des évolutions plus lentes comme la croissance de cultures agricoles, le développement d'algues dans les rivières ou l'évolution de couvertures forestières. Enfin, on observe aussi des structures stables dans les zones urbaines.
L'échelle spatiale à laquelle interviennent les phénomènes est elle aussi variable. Par exemple, le survol de la scène par un avion sera caractérisé par un nombre faible de pixels comparé au nombre de ceux impliqués dans la caractérisation d'un phénomène comme la moisson.
Enfin, les structures spatio-temporelles sont caractérisées par des plages de radiométries

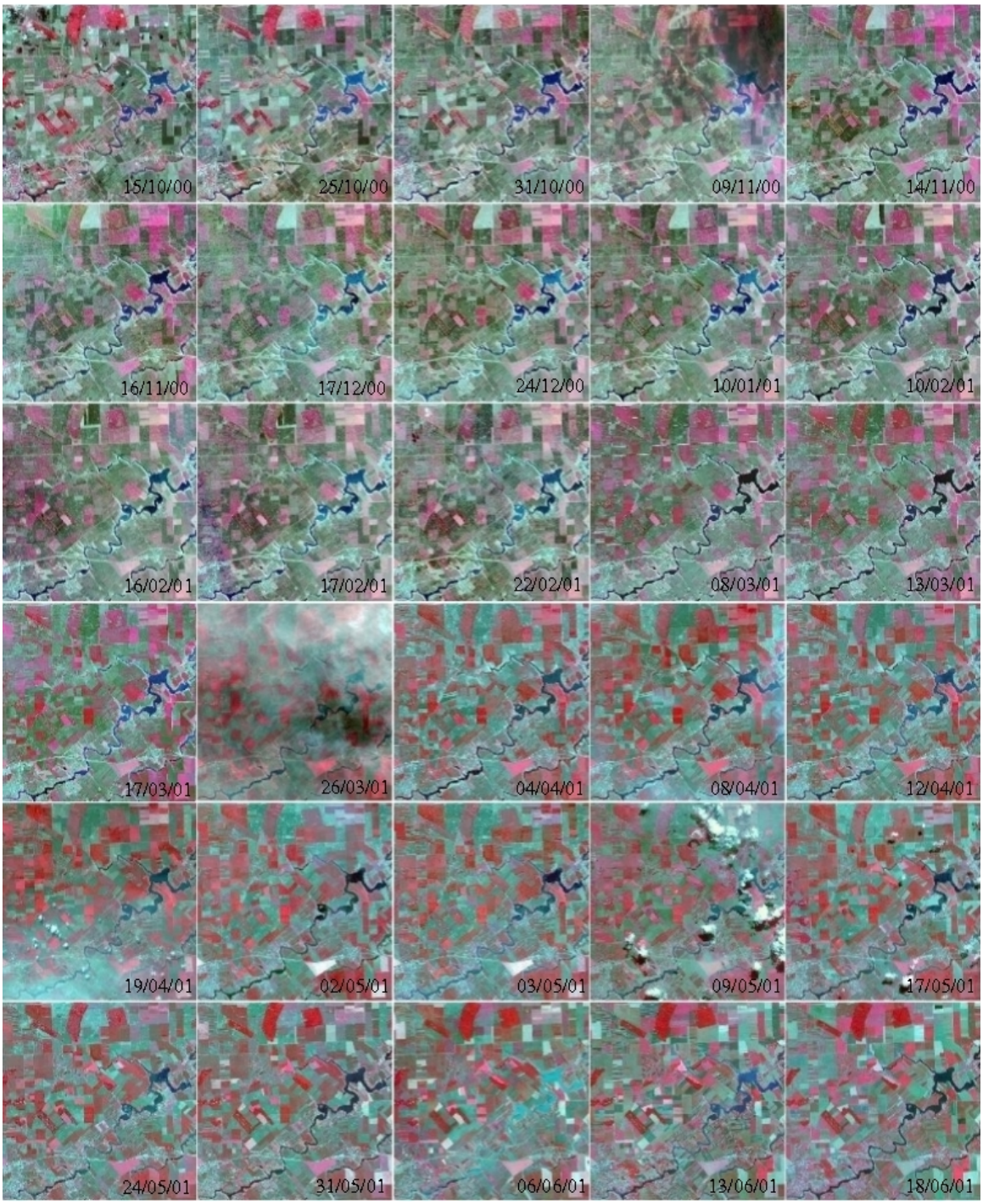


FIG. 2.2 – *Fenêtre spatiale de la STIS ADAM prétraitée représentant l'évolution de la zone rurale de Fundulea (Roumanie). Les images présentées contiennent 800x800 pixels (256km^2), et ont été acquises entre le 15 octobre 2000 et le 18 juin 2001. Pour la visualisation, chacune des images de la série est sous échantillonnée et possède un histogramme normalisé (c.f.* § *5.2).*

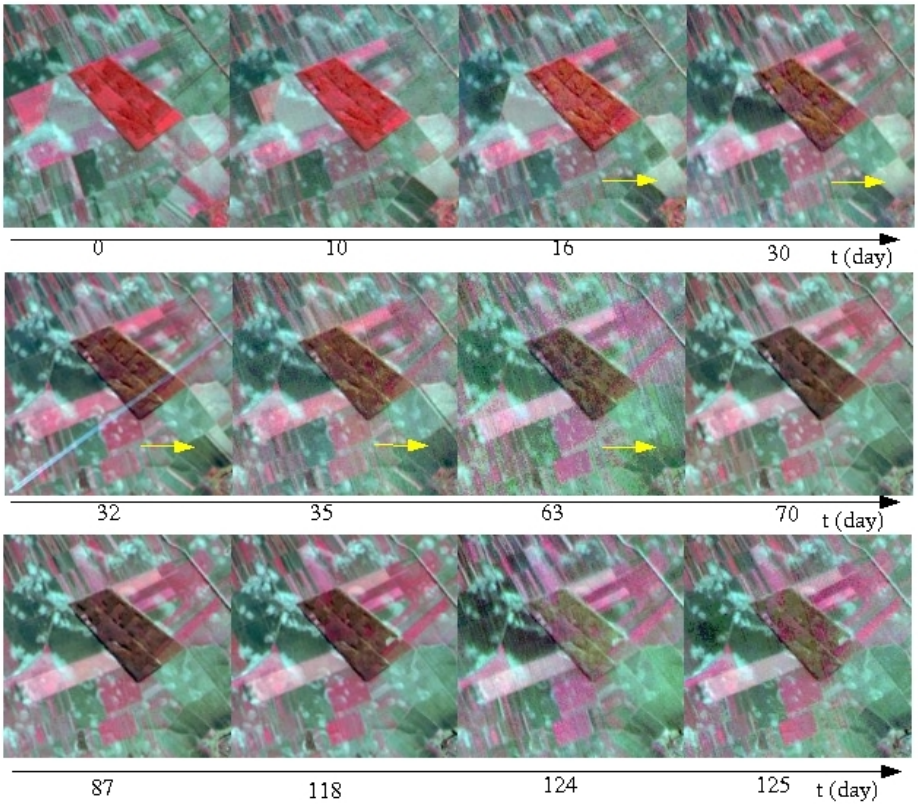


FIG. 2.3 – *De nombreuses structures spatio-temporelles sont présentes dans les STIS. Par exemple, dans la STIS ADAM, les phénomènes de moissons et les labours sont nombreux. Une flèche signale l'un d'entre eux. D'autre part, on observe l'apparition d'un avion dans l'image acquise le* 32*-ième jour (14 novembre 2001), ou encore l'évolution progressive d'une couverture forestière.*

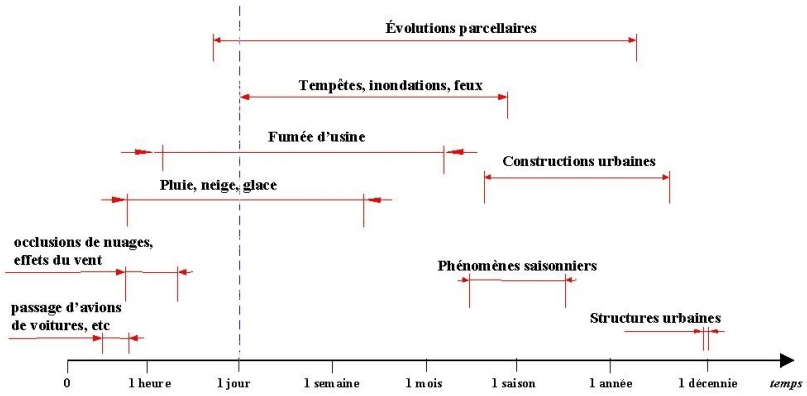


FIG. 2.4 – *Les processus de changements liés aux phénomènes intervenant dans une STIS ont des échelles temporelles différentes. Par exemple, les apparitions d'avions constituent des structures évanescentes caractérisées par des échelles temporelles courtes tandis que les bâtiments sont des structures stables caractérisées par des échelles temporelles relativement longues. Les processus de changements liés aux évolutions agricoles, tels que la croissance des plantes ou les phénomènes de moisson, sont à des échelles temporelles variées. Les structures caractérisées par des échelles temporelles en dessous de la ligne pointillée bleue possèdent généralement dans leur spectre des fréquences plus hautes que la fréquence théorique de la limite d'échantillonnage.*

variées. Par exemple, la variation radiométrique engendrée par la floraison d'une parcelle de colza est beaucoup plus importante que celle induite par la maturation d'un champ de blé.
En conséquence, afin de caractériser efficacement ces différentes structures possédant des échelles spatiales et temporelles hétérogènes ainsi que des plages radiométriques différentes, les méthodes d'analyses devront se situer du niveau pixellaire jusqu'au niveau d'une structure spatiale importante (ville, forêt, ...) afin de modéliser des évolutions des radiométries rapides ou lentes, et de faible ou de grande amplitude.

Cependant, le jeu de données ADAM, et d'une façon plus générale les STIS de haute résolution, imposent plusieurs contraintes d'analyses.
D'abord, malgré les améliorations notoires effectuées pour la calibration radiométrique par un ajustement absolu puis relatif des réflectances, il existe toujours des discontinuités temporelles des radiométries. Ces "sauts de réflectances" perturbent la modélisation temporelle des structures.
Ensuite, l'analyse temporelle de la STIS au niveau pixellaire est difficile car

1) l'on dispose d'un petit nombre d'échantillons (38 images), ce qui rend difficile toute modélisation stochastique,
2) les contraintes telles que l'apparition de nuages ou la disponibilité du capteur empêchent l'acquisition régulière de la même scène. Les processus temporels sont donc irrégulièrement échantillonnés,
3) les difficultés de superposabilité des images sont accrues par la haute résolution et mènent à des structures spatiales légèrement mobiles dans le temps pouvant perturber

une analyse à un niveau pixellaire,

4) la fréquence d'échantillonnage temporel de beaucoup de motifs accessibles grâce à la haute résolution, transgresse le théorème de l'échantillonnage ; en effet, beaucoup de structures spatio-temporelles possèdent dans leurs spectres des fréquences plus hautes que la fréquence théorique limite d'échantillonnage ; or ces structures sont décrites par seulement quelques échantillons temporels. Celles-ci sont donc sous-échantilonnées temporellement.

Concernant ce dernier point, notons que les satellites acquièrent généralement une seule image du même site par jour. En conséquence, l'on peut distinguer 2 catégories de structures spatio-temporelles : des structures présentes dans une seule image et des structures dont on peut observer l'évolution. La démarcation entre ces deux types de structures est représentée dans la figure 2.4 par la ligne verticale pointillée.

Face à ces contraintes, une analyse à l'échelle spatiale des structures sous échantillonnées pourra améliorer la description de sa dynamique et autoriser une investigation en dessous de la fréquence limite théorique. Par exemple, le panache de fumée d'une usine, qui constitue une structure évoluant dans le temps du fait de l'activité de l'usine, et dans l'espace du fait de la directionnalité du vent, pourra être modélisée d'une façon plus pertinente en utilisant une analyse de la structure "fumée" présente dans toute l'image, qu'en effectuant une analyse des séries temporelles au niveau pixellaire.
Par ailleurs, une telle approche contextuelle pourra améliorer l'analyse des structures en utilisant leur complémentarité. Par exemple, l'étude du comportement radiométrique d'un pixel d'un champ dont l'observation a été détériorée par l'apparition d'aérosols, sera plus robuste par la prise en compte du comportement de pixels de champs similaires non perturbés. De plus, une analyse par régions sera plus robuste face aux problèmes de superposabilité.

2.4 Espaces de représentations

Les STIS sont des processus stochastiques multidimensionnels compris dans un espace formé à partir de deux axes spatiaux r, d'un axe temporel t et de plusieurs composantes de caractéristiques spectrales θ. Dans cet espace on représente le signal, noté $x(r, t, \theta)$, sous la forme d'un nuage de points. Pour modéliser ce signal, l'on doit analyser conjointement l'évolution spatiale et l'évolution temporelle des caractéristiques.

Beaucoup de méthodes sont disponibles dans la littérature pour analyser l'évolution spatiale des caractéristiques. Une image de la série est un processus non-stationnaire, régulièrement échantillonné et caractérisé par un nombre important de réalisations. L'analyse locale de ce processus spatial permet la génération de nouvelles composantes de caractéristiques de texture, de forme, etc. Ces nouvelles caractéristiques peuvent être ajoutées aux caractéristiques spectrales, pour former un nouveau vecteur de caractéristiques θ.
De nombreuses méthodes existent pour l'analyse de l'évolution temporelle des caractéristiques au niveau pixellaire. Mais dans le cas des STIS, une série temporelle localisée au

niveau d'un pixel est un processus non-stationnaire, caractérisé par un faible nombre de réalisations, irrégulièrement échantillonné. Les difficultés rencontrées pour effectuer une telle analyse ont été exposées dans la section précédente.

Pour contourner ces dernières difficultés et pour modéliser conjointement l'évolution spatiale et l'évolution temporelle des caractéristiques, on s'intéresse plutôt à l'analyse contextuelle et temporellement localisée du processus. On recherche donc des représentations du signal $x(r,t,\theta)$ amenant adaptées à une telle analyse.

Espace naturel L'espace naturel de représentation de la STIS est noté $\mathcal{E}_{r,t,\theta}$. Dans cet espace on représente la STIS originale, c'est à dire le nuage de points $x(r,t,\theta)$. Ce dernier contient l'information sur l'évolution des caractéristiques dans l'espace euclidien et dans le temps.
Notons que dans cet espace naturel de représentation, pour rehausser la qualité visuelle d'une certaine information, on peut représenter la STIS avec des radiométries modifiées.

Espace des caractéristiques évolutives L'espace formé à partir des composantes de caractéristiques θ et de la composante temporelle t est noté $\mathcal{E}_{t,\theta}$. Il est appelé espace des caractéristiques évolutives. Dans cet espace on représente la STIS sous la forme d'un histogramme de caractéristiques multidimensionnelles évoluant dans le temps. Les caractéristiques évolutives dépendent de localisations spatiales r. On note donc ce signal $x_r(t,\theta)$. Celui-ci contient l'information sur l'évolution de la distribution des caractéristiques dans le temps ; l'information spatio-temporelle du signal est cachée mais reste accessible si l'index spatial r de chaque caractéristique dans cet espace est sauvé.

Espace spatio-temporel L'espace spatio-temporel est formé à partir des composantes spatiales r et de la composante temporelle t. Il est noté $\mathcal{E}_{r,t}$. En parallèle et complémentairement aux représentations dans l'espace des caractéristiques évolutives, on peut représenter dans cet espace l'évolution spatiale et temporelle associée à certaine structure sous la forme d'un nuage de points. On note ce signal $x(r,t)$. Ensemble, les représentations dans l'espace spatio-temporel et dans l'espace des caractéristiques évolutives, caractérisent le signal $x(r,t,\theta)$.

Espaces Temporellement Localisés (TL) des caractéristiques Notons par d le nombre d'échantillons temporels de la série d'images et notons par $\{\theta_{t_i}; i = 1,...,d\}$ les composantes de caractéristiques TL. On forme d espaces à partir des d différentes composantes de caractéristiques θ_{t_i} localisées en t_i. On note chaque espace par $\mathcal{E}_{\boldsymbol{\theta}_i}$. Les d espaces formés sont appelés espaces TL des caractéristiques. Dans ces espaces on représente la STIS sous la forme d'une succession d'histogrammes de caractéristiques multidimensionnelles. Ces caractéristiques dépendent de la localisation spatiale r. On note donc ces signaux $\{x_r(\theta_{t_i}); i = 1,...,d\}$. Chaque signal $x_r(\theta_{t_i})$ contient l'information sur la distribution des caractéristiques TL en t_i.
On remarque que ces espaces temporellement localisés des caractéristiques constituent les

différents états de l'espace des caractéristiques évolutives.
Par ailleurs, si l'on regroupe les caractéristiques par similarité, alors on obtient pour chaque localisation t_i une collection de K_{t_i} clusters notée $\{C_{t_i}^j; j = 1, ..., K_{t_i}\}$. Ces derniers sont bien sûr représentés dans l'espace TL des caractéristiques $\mathcal{E}_{\theta_{t_i}}$.

Espaces des classes TL Chaque classification spatiale $\{C_{t_i}^j(r); j = 1, ..., K_{t_i}\}$ associée à une collection de clusters TL $\{C_{t_i}^j; j = 1, ..., K_{t_i}\}$ est représentée dans l'espace spatial noté $\mathcal{E}_r$. On obtient donc d classifications différentes qui sont notées $\{x_{t_i}(r);\ i = 1, ..., d\}$. Ces signaux contiennent l'information spatiale relative aux clusters TL.
On remarque que les espaces des classes TL constituent les différents états des espaces spatio-temporels associés aux clusters dynamiques.

Espace MultiTemporel (MT) des caractéristiques L'espace composé par l'ensemble des dimensions de caractéristiques temporellement localisées $\{\theta_{t_i}\}$ aux différents instants $\{t_i\}$ avec $i = 1, ..., d$ est noté $\mathcal{E}_{\theta_{t_1},...,\theta_{t_d}}$. Cet espace est appelé espace MT des caractéristiques. Dans cet espace, on représente la STIS sous la forme d'un histogramme de caractéristiques multidimensionnelles. Ces caractéristiques dépendent de la localisation spatiale r. On note donc ce signal $x_r(\theta_{t_1}, ..., \theta_{t_d})$. Cet espace contient comme l'espace des caractéristiques évolutives, l'information sur l'évolution de la distribution des caractéristiques. Mais dans cette représentation du signal, les composantes temporelles ont perdu leur ordre chronologique[3].
Notons qu'en projetant l'espace MT dans des espaces temporellement localisés des caractéristiques, on peut décomposer cet espace MT des caractéristiques et reconstituer les différents états de l'espace des caractéristiques évolutives.
Par ailleurs, si l'on regroupe les caractéristiques par similarité, alors on obtient une collection de K_{MT} clusters notée $\{C_{MT}^j; j = 1, ..., K_{MT}\}$. Ces derniers sont bien sûr représentés dans l'espace MT des caractéristiques.

Espaces des classes MT Chacune des classes spatiales $\{C_{MT}^j(r); j = 1, ..., K_{MT}\}$ associées à la collection de clusters MT $\{C_{MT}^j; j = 1, ..., K_{MT}\}$ est représentée dans l'espace spatial $\mathcal{E}_r$. On obtient donc une classification notée $x_{MT}(r)$. Ce signal contient l'information spatiale relative aux clusters MT.

[3]Une alternative à la représentation du signal dans l'espace MT est la constitution d'un nouvel espace, noté $\mathcal{E}_{\theta(t)}$. Pour le constituer, on procède à l'extraction de caractéristiques $\theta(t)$ au niveau de chaque série temporelle localisée en chaque pixel de l'image. Dans cet espace formé à partir des composantes $\theta(t)$, on représente la STIS sous la forme d'un histogramme noté $x_r(\theta(t))$. Celui-ci contient l'information sur l'évolution temporelle globale des caractéristiques au niveau de chaque pixel. En conséquence, la chronologie temporelle est respectée, mais la décomposition temporelle et la reconstitution des différents états de l'espace des caractéristiques évolutives n'est plus possibles.
De plus, comme nous l'avons vu dans la section précédente, pour obtenir cette représentation du signal, nous sommes confrontés à l'analyse délicate des séries temporelles localisées en chaque pixel.

2.5 Résumé

Dans ce chapitre, les points suivants ont été détaillés.

- La série temporelle d'images SPOT, utilisée dans ces travaux de thèse, a d'abord été présentée. Le processus de constitution de la STIS a par ailleurs été détaillé.

- Nous avons ensuite vu que les structures spatio-temporelles des STIS représentent une énorme quantité d'information et nous avons décrit les particularités de telles structures. Après avoir exposé des généralités sur l'analyse de telles structures, nous avons introduit les problèmes d'analyse spécifiques aux structures spatio-temporelles des STIS. En particulier, nous avons souligné que l'échantillonnage irrégulier, le nombre insuffisant d'échantillons temporels et les difficultés de corrections radiométriques/géométriques rendent l'analyse temporelle des motifs difficile et qu'une analyse par régions paraît être une approche plus adaptée qu'une analyse par pixel.

- Finalement, nous avons introduit des espaces de représentation de ces signaux multidimensionnels utiles à une analyse contextuelle et temporellement localisée. L'information présente dans ces espaces porte sur l'espace euclidien, le temps et les caractéristiques. Les relations entre les différents espaces de représentation ont été par ailleurs établies. En particulier, nous avons vu que l'espace naturel peut engendrer des espaces complémentaires : l'espace des caractéristiques évolutives, qui exhibe l'information sur l'évolution de la distribution des caractéristiques, et l'espace spatio-temporel qui exhibe l'information sur l'évolution spatiale des structures. Il a été aussi remarqué que l'espace multitemporel des caractéristiques, contenant l'information sur l'évolution de la distribution des caractéristiques, se décompose en sous-espaces appelés espaces temporellement localisés des caractéristiques, qui constituent les différents états de l'espace des caractéristiques évolutives.

Chapitre 3

La fouille d'information

3.1 Motivations

L'information peut être définie comme les connaissances acquises par l'étude du contenu d'un système. Comme le souligne *Jumarie (1990)*, l'information est relative. Par exemple, l'information contenue dans un "magazine λ" n'est pas la même pour un homme d'affaire ou pour un demandeur d'emploi. De même, l'information contenue dans l'ADN est relative à son environnement. Si l'on place l'ADN dans un environnement étranger, alors l'information qu'il contient n'est plus exploitable et devient inutile.

L'information inonde le monde actuel. Elle est présente dans les messages transmis par les moyens standards de communication tels que la radio, la télévision ou Internet, mais aussi dans les messages véhiculés au travers de données bancaires, les signaux impliqués dans l'électronique des ordinateurs, dans l'imagerie médicale ou satellitaire, ou encore dans les signaux transitant dans le système nerveux de l'animal ou de l'Homme. Du fait d'importants progrès technologiques, la quantité d'information disponible augmente considérablement et présente un intérêt potentiel croissant pour la prise de décision. Une des raisons de l'accroissement incessant de données est l'automatisation d'activités dans de nombreux domaines tels que les affaires, l'ingénierie ou les sciences. De nos jours, une simple transaction telle que l'utilisation d'une carte de crédit ou d'un téléphone peut être enregistrée pour constituer une base de donnée électronique au potentiel d'information gigantesque. Parmi tous ces types de données, l'un d'entre eux prend une place de plus en plus importante dans notre société : l'image numérique. Les images contiennent des millions de pixels, chacun porteur d'un élément d'information. En particulier, concernant les systèmes satellitaires d'observation, ces éléments d'information se multiplient jusqu'à former quotidiennement plusieurs Tera-Octect de données.

Les technologies actuelles de recherche de documents supportent des grandes masses de

données : d'après un article de presse mis à disposition par Google Inc. en mars 2003, le moteur de recherche permet l'accès à plus de 3 milliards de documents Web, servant plus de 200 millions de recherches par jours et qui retourne plus de 330 millions images. AltaVista supporte plus de 25 millions de recherches par jour dans plus de 25 langues, avec sa partie de recherche multimédia qui retourne plus de 240 millions d'images, vidéos et audio.

Cependant, les méthodes actuelles pour accéder à l'information dans ces documents sont rudimentaires. Actuellement les technologies développées pour la recherche de documents utilisent principalement des techniques basées sur des mots-clés associés aux documents pour représenter leurs contenus informatifs ; cette manière d'indexer est fortement sous-optimal autant sur le plan de la consommation de ressources humaines pour l'annotation manuelle des documents, que sur le surplus de stockage nécessaire ou la non-objectivité de la description.
Concernant l'exploitation d'images satellites, actuellement, les systèmes opérationnels autorisent des recherches par localisations géographiques, par date d'acquisition ou par type de capteur. Cette information est souvent une description sommaire du contenu de la scène.

Le besoin de transformer les grandissantes bases de données souvent complexes et multidimensionnelles, en information et connaissances devient une priorité dans l'industrie de l'information. En effet, la vitesse d'acquisition de ces données dépassant la capacité humaine pour les interpréter manuellement, on est confronté à un accroissement d'une masse de données non exploitée.
La fouille d'information s'attache à résoudre cette problématique, conciliant l'Homme et la Machine, en adaptant l'information contenue dans les données à la conjecture de l'utilisateur.

3.2 Définition de la fouille d'information

3.2.1 Le paradigme de la fouille d'information

La fouille d'information est le processus engagé pour explorer et découvrir des connaissances, tels que des motifs, des associations, des règles, des relations de cause à effet, des changements, des anomalies et des structures significatives, à partir de grandes quantités d'information stockées dans des bases de données ou autres "mines" d'information. La figure 3.1 illustre ce paradigme. La fouille d'information peut encore être définie comme le processus non trivial d'analyser des données dans l'optique de trouver une information implicite et potentiellement utile. Elle permet une recherche par le contenu et amène à la découverte de connaissances.

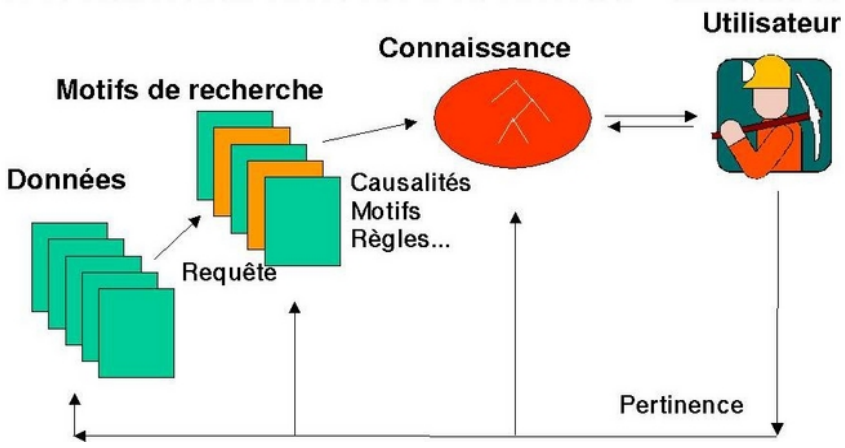


FIG. 3.1 – *Le paradigme de la fouille d'information.*

Par ailleurs, une définition mathématique de la fouille d'information peut être formulée *(Keim et Kreigel, 1996)*. On considère pour l'analyse, le jeu de données D. La fouille d'information peut être décrite comme le processus consistant à trouver :

- un sous-ensemble $D^{'}$ de D et
- des hypothèses $H_U(D^{'}, C)$ vérifiées par $D^{'}$

qu'un utilisateur U considère intéressantes dans le contexte C d'une application. La discrimination du sous-ensemble $D^{'}$ constitue le résultat visé dans le concept de recherche par le contenu. Les hypothèses sont la formulation de la conjecture, floue ou nettement définie, de l'utilisateur. Elles expriment l'intérêt d'un utilisateur et peuvent être des propriétés de similarité vérifiées par tout le sous-ensemble de données $D^{'}$. Elles peuvent aussi être des dépendances fonctionnelles exprimant des contiguïtés ou des relations de cause à effet. Ces hypothèses constituent des modèles de connaissances, découvertes ou vérifiées par la fouille. Les éléments du sous-ensemble $D^{'}$ sont appelés motifs relatifs à $H_U(D^{'}, C)$.

Ce processus est généralement itératif et comprend l'interaction d'un utilisateur dans la boucle. L'utilisateur commence avec une hypothèse et utilise les données pour la réfuter ou la confirmer. L'hypothèse est affinée, en fonction de la réponse et ce processus est poursuivi jusqu'à ce qu'un stade satisfaisant de connaissances ait été obtenu. L'intervention de l'utilisateur dans le système se fait, sous la forme d'apport de connaissances liées au domaine et peut s'inscrire dans le contexte d'un apprentissage supervisé, semi-automatique ou automatique.

3.2.2 Complémentarité de l'Homme et de la machine pour la fouille

Les paragraphes suivants s'attachent d'abord à exposer les mécanismes de la cognition et les caractéristiques du système perceptif humain, pour ensuite dégager l'intérêt de l'intégration d'un apprentissage semi-automatique dans le processus de fouille. On se limitera au cadre de la vision.

La cognition et la perception visuelle humaine

Le processus engagé par l'Homme lors de sa confrontation à une source d'information constitue un vaste sujet d'étude. Des tentatives visant à modéliser les mécanismes humains du traitement de l'information ont conduit à des modèles de fonctionnement de la cognition humaine *(Wiener, 1961 ; Niculiu, 2001)*. Dans le domaine de la vision, *Eakins (1996)* décomposa le processus de traitement de l'information en niveaux hiérarchiques : la perception humaine d'une information, c'est à dire la collecte par la vue de caractéristiques ou primitives (formes, textures, couleurs, ...), la représentation sémantique (chaise, arbres, neige, ...) et la compréhension (phénomène de pollution, émotion de tristesse, ...). Bien que ce modèle hiérarchique de l'information s'apparente à un modèle plus cognitif que seulement perceptif, il l'appela le modèle gradient de la perception visuelle humaine.

L'Homme possède une capacité remarquable pour le traitement de données incomplètes ou bruitées. En effet, pour traiter de telles données, l'information a priori est largement utilisée dans le mécanisme d'interprétation. Elle est propre à chaque individu et dépend du contexte dans lequel intervient l'information. Les principes de la *Gestalt* démontrent la prise en compte de la nature organisée de l'information environnante, constituant des a priori fort, pour l'inférence de raisonnements *(Koffka, 1935)*.

Cette aptitude à recevoir et traiter pertinemment l'information est cependant contrainte par les capacités limitées du système de perception visuelle. Dans le domaine de la psychologie perceptuelle, plusieurs expérimentations ont été conduites pour étudier ces limitations.

Une contrainte évidente est l'incapacité du système perceptif humain à analyser de trop important volume de données car sa capacité de mémorisation est limitée.

La capacité humaine limitée pour mesurer de façon absolue l'information visuelle constitue une autre contrainte. En modélisant la perception humaine comme un canal de communication, des expérimentations ont montré que la quantité d'information absolue transmise en bits ne constitue qu'une faible proportion de l'information reçue *(Miller, 1956)*. Par contre, la perception absolue d'une information multidimensionnelle s'effectue avec une meilleure acuité, ce qui conforte l'idée soutenue par les linguistes que l'Homme utilise peu d'information sur beaucoup de paramètres. Par contre, la capacité humaine à percevoir des informations relatives est bien plus importantes *(Cleveland, 1993)*.

Une dimensionnalité trop élevée des données est contraignante. Premièrement, des données décrites dans un espace avec un nombre de dimensions spatiales supérieures à 3 ne sont pas directement visualisables. Deuxièmement, l'incorporation dans cet espace de la dimension temporelle implique la visualisation d'un espace dynamique ; cette tâche de visualisation est parfois difficile pour le système perceptif humain car elle fait appel à la mémorisation d'évolutions passées. Enfin, trop de dimensions supplémentaires comme la couleur, la texture, etc, ne pourront pas être perçues simultanément par le système de perception humaine. Toutefois, pour l'exploration d'espaces multidimensionnels, des techniques de visualisation sont disponibles (c.f. § 3.3.5) ; mais elles engendrent dans la plupart des cas une perte d'information.

Aptitudes des machines et intelligence artificielle
Une machine peut effectuer avec une vitesse supérieure et de façon précise certaines tâches que l'Homme juge difficile.
La machine mesure de façon absolue l'information, tâche difficilement exécutable pour l'Homme.
De plus, la machine peut mesurer et traiter l'information contenue dans un signal de dimension très élevée. Pour la visualisation, des représentations pertinentes pour l'Homme, donc de plus faible dimension, pourront être recherchées.
D'autre part, la capacité de mémorisation d'information absolue d'une machine est nettement supérieure à celle de l'Homme. Cette aptitude autorise l'analyse de grande masse d'information.

L'intelligence d'un système est définie comme ses capacités à construire des modèles de représentation du monde réel, d'apprendre et d'effectuer des inférences. La complexité de la cognition humaine rend la simulation de l'intelligence humaine difficile. En conséquence des raisonnements de haut niveau ont du mal à être modélisés. Toutefois, des modèles sémantiques ont été développés ces dernières années. Ils aident l'Homme dans les processus d'interprétation. Dans les chapitres suivants, un état de l'art présentera ces modèles (c.f. § 3.3).

Complémentarité
La fouille peut s'effectuer en exploitant les capacités de l'Homme seul, celles d'une "machine intelligente" ou en couplant les deux.
Dans la première situation, en fonction des caractéristiques des données, l'Homme se confrontera plus ou moins rapidement à ses limites de perception et d'analyse.
Dans la deuxième situation, la définition de fouille d'information peut s'effectuer en préformulant un jeu d'hypothèses et en recherchant l'information vérifiant ces hypothèses. Une bonne formulation des hypothèses aboutira à des résultats corrects, obtenus de façon non-supervisé. Cependant, la complexité du système cognitif humain peut rendre la formulation d'hypothèses relatives à la conjecture d'un utilisateur difficile. De plus, de tels systèmes seront dédiés à des applications prédéfinies et ne seront pas flexibles.
Dans la dernière situation, les aptitudes complémentaires de l'Homme et de la machine seront exploitées pour obtenir l'information pertinente. Par ailleurs, la fouille adoptera une optique exploratoire, sans hypothèses préformulées, capable de s'adapter à presque n'importe quelle application et notamment aux nouveaux scénarios engendrés par des avancées scientifiques.

3.2.3 La fouille d'information vue comme un problème de canal de communication

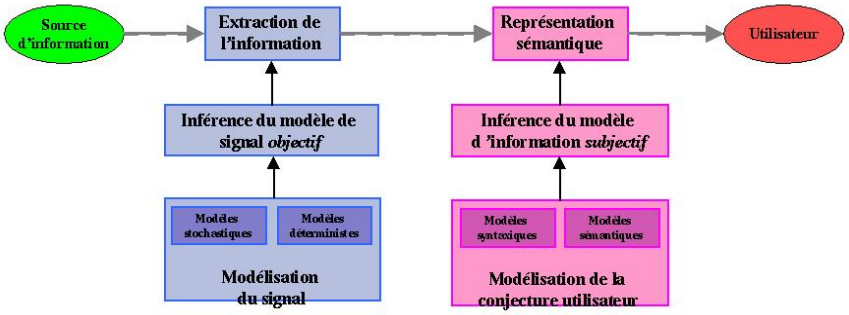


FIG. 3.2 – *La fouille d'information est un concept de communication.*

Le problème de la fouille peut se comprendre comme un problème de canal de communication entre d'un coté une base de données source d'information, et de l'autre un utilisateur jouant le rôle de récepteur *(Datcu et Seidel, 1999a ; Datcu, 1999b)*. La figure 3.2 illustre ce concept. Ce concept est une décomposition du processus de traitement de l'information en 2 niveaux hiérarchiques correspondant aux 2 premiers niveaux du modèle gradient de la perception visuelle humaine d'*Eakins (1996)*.

Pour construire ce canal de communication, on l'adapte d'un côté aux données et de l'autre aux sémantiques des utilisateurs. Le long du canal, on représente l'information sous forme hiérarchique :

- **Représentation "signal"**. Cette représentation est obtenue en capturant l'information par modélisation déterministe ou stochastique du signal. L'information extraite aux travers de ces modèles, est représentée dans un langage propre à la modélisation effectuée, sans signification pour l'utilisateur.
  Généralement deux traitements de l'information sont impliqués pour engendrer la représentation "signal" :
  – des caractéristiques sont extraites au travers d'une modélisation des données et représentées dans un espace particulier ; celles-ci doivent être concises, les plus complètes possibles et avoir un fort pouvoir discriminant ;
  – les caractéristiques sont ensuite soumises à des analyses basées sur des modèles dans lesquelles l'information pertinente est "filtrée" et des mesures de similarité sont définies ; à ce niveau, des représentations pertinentes de l'information sont recherchées, des similarités sont dégagées et l'information "filtrée" est indexée d'une façon appropriée.
  On note que dans de nombreux domaines, il existe une dégradation de l'information réelle acquise par les capteurs. Par exemple, des contraintes imposées par le capteur sur la résolution spatiale ou sur la quantité de bruit, engendrent une image dégradée de la scène réelle. Cette perte d'information se nomme le "fossé du capteur".

- **Représentation "sémantique"**. Cette représentation est obtenue par modélisation de la conjecture de l'utilisateur. L'information est alors représentée par l'intermédiaire de modèles sémantiques et syntaxiques compréhensibles par un utilisateur.
  Pour induire cette représentation, les modèles sémantiques sont ajustés lors par apprentissage supervisé. Pour guider l'apprentissage, des visualisations pertinentes de l'information sont recherchées et des mesures de pertinence peuvent être développées. La recherche par le contenu, la découverte de connaissances et la compréhension des données sont dès lors accessibles.
  Très souvent, il existe un écart entre l'information extraite par la modélisation des données et l'information que l'utilisateur peut extraire directement des données non-modélisées. Par exemple, il existe un fossé entre les caractéristiques extraites d'une image (i.e. forme, texture, couleur, etc) et les sémantiques qu'un utilisateur peut associer à cette même image (plaque de mazout, pollution marine, etc). Cette différence se nomme le "'fossé sémantique".

La fouille d'information est un domaine jeune et interdisciplinaire basé sur les statistiques, la théorie de l'information et l'apprentissage automatique mais aussi fortement lié aux techniques de visualisation de l'information, aux systèmes de base de données et à l'informatique en général. De plus, il implique la compréhension du fonctionnement de la cognition humaine.

3.3 Etat de l'art de la fouille

3.3.1 La fouille de données

Un intérêt particulier a été porté pour la gestion et la manipulation de données bancaires, commerciales ou plus généralement issues du domaine des affaires.
Il existe une multitude de données bancaires, commerciales, de consommation, textes, etc. La mine d'information constituée par ces données nécessite des moyens efficaces d'exploitation et d'interprétation. La fouille d'information ouvre de nouvelles perspectives.

Modèle sémantique de données bancaires. Un système capable de découvrir, à partir d'une base de données particulière, des règles propres qui la régissent a été développé par *Agrawal (1993)*. Les règles envisageables découlent de modèles et formalisent des hypothèses envisageables. Les règles sont constituées par une combinaison d'opérations élémentaires. La fouille automatique s'intéresse uniquement aux règles satisfaisant certaines contraintes syntaxiques et de support, dictées par l'utilisateur. Les contraintes syntaxiques impliquent des restrictions sur la composition des règles et les contraintes de support concernent un nombre minimal de cas requis supportant la règle. Ce formalisme permet la résolution de divers problèmes de fouille de données tels que la classification de données pour une étude de marché, la découverte de règles d'associations modélisant les habitudes de consommation d'une clientèle, ou l'établissement de règles régissant des séquences temporelles afin de comprendre et prédire l'interaction entre les différents cours

d'actions bancaires.
Pour la fouille de données financières ou de nature semblable, les règles développées constituent des modèles sémantiques pertinents. Mais les méthodes développées sont spécifiques au contexte et limités dans le choix de modèles.

Indexation sémantique de données textuelles. Presque tout utilisateur d'Internet emploie des moteurs de recherches comme Google, qui aide à retrouver des documents avec un contenu particulier. La recherche est basée sur le contenu dans le sens qu'un index contenant tous les mots de chaque document existe et autorise le système à retrouver efficacement les documents semblables. En d'autres termes, pour des documents textuels, la sémantique du document est assimilée aux données textuelles elles même.
Mais la recherche basée sur l'indexation de mots ne permet pas la compréhension du document. D'autres stratégies telles que l'indexation sémantique latente, utilisent des modèles statistiques de classes latentes qui permettent des représentations sémantiques des mots et des documents *(Hofmann, 1999 ; Deerwester et al., 1990)*.

3.3.2 La fouille d'information dans les images

Les systèmes d'information relatifs aux images doivent être doués de fonctionnalités issues à la fois de la gestion des bases de données et de la théorie de la vision par ordinateur ; mais il existe un fossé entre ces deux systèmes. La théorie des bases de données n'étant pas adaptée pour manipuler des structures picturales multidimensionnelles et les systèmes de vision n'offrant pas de possibilités pour effectuer des recherches. Pour couvrir une si large problématique, une approche unificatrice doit être adoptée où les données de l'image sont considérées comme une source d'information et où l'utilisateur est perçu comme un récepteur. En conséquence, l'accent doit être mis sur la modélisation du contenu de l'information de l'image, la conjecture de l'utilisateur, l'organisation de l'information et l'interaction Homme-Machine.

La recherche d'image par le contenu, la découverte de connaissances et la compréhension de scène partagent des méthodes communes (analyse de données et techniques de modélisation probabiliste) et des objectifs identiques (l'extraction d'information utile à partir de données affectées par des incertitudes). Par contre, elles opèrent à des échelles de quantité de données différentes.
La *recherche d'image par le contenu* est un processus de fouille d'images permettant l'identification de structures dans les images, qui possèdent des tailles conséquentes ou qui sont fortement caractérisées. Un espace interne limité dans la base de donnée est nécessaire pour chaque image autorisant l'application de la méthode à des archives entières d'images. Les applications d'intérêt sont par exemple, l'indexation par le contenu d'importantes bases de données d'images.
Un intérêt particulier est également porté sur la *découverte de connaissances* et la *compréhension de scène* dans les images. La découverte de connaissances est la fouille d'information permettant l'identification de structures de taille spatiale moyenne et des règles les régissant basées sur leurs similarités, leurs interactions spatiales, etc. Un espace

relativement important est nécessaire pour le stockage dans la base de données de chaque image caractérisée, et en conséquence, la fouille est applicable à des jeux d'images de taille limitée. La compréhension de scène permet l'identification d'information et de structures détaillées, à une résolution pixellaire et en conséquence la compréhension détaillée de la scène. L'espace que requiert la caractérisation de chaque image est encore plus important, limitant ainsi la fouille à seulement quelques scènes.

Il existe beaucoup d'approches pour l'extraction d'information des images et la modélisation de leurs sémantiques. Quelques-unes des plus efficaces sont brièvement décrites dans la suite.

QBIC (Query By Image Content) a été le premier système commercial de fouille d'image par le contenu *(Flickner et al., 1995)*. Son architecture et ses techniques ont profondément influencé beaucoup de systèmes actuels. Il supporte des recherches basées sur des exemples d'images ou sur des motifs de couleurs ou textures désignés par un utilisateur. Des caractéristiques de couleur, texture et de forme sont extraites de chaque image. Les primitives multidimensionnelles subissent une réduction de dimensionnalité et sont ensuite indexées efficacement par l'utilisation d'une structure $R^* - tree$ *(Beckmann et al., 1990)*. Le modèle sémantique du système est basé sur des similarités de caractéristiques codées dans cette structure de $R^* - tree$. Ce modèle permet une indexation efficace mais ne permet pas une modélisation sémantique interactive.

Li et Wang (2003) décrivent des catégories d'images, chacune correspondant à un concept, par des modèles stochastiques. En particulier un modèle de Markov caché, combiné avec une décomposition en ondelettes de la couleur et de la texture, est utilisé pour extraire l'information représentative d'une catégorie d'images. Puis, une indexation automatique des images est effectuée en les associant avec des termes linguistiques. Celle-ci compare statistiquement l'image à indexer avec la collection de modèles d'apprentissage formant le dictionnaire de concept.
Cette approche a pour avantage de tenir compte des relations spatiales dans l'image via le modèle statistique. De plus, l'association avec les termes linguistiques permet une représentation sémantique basée sur des mots.

Un modèle statistique pour organiser une collection d'images selon la sémantique d'un utilisateur a été proposé par *Barnard et Forsyth (2001)*. Chaque image est modélisée comme étant générée par une hiérarchie fixée de noeuds, où les noeuds de l'arborescence correspondent à des clusters. En effet, chaque noeud de l'arbre possède une probabilité de générer un mot et une probabilité de générer un segment d'image (groupe de pixels) par l'intermédiaire de caractéristiques. Les images sont modélisées comme étant générées par les noeuds appartenant à un chemin particulier dans l'arbre.
L'information sémantique est issue de l'association de texte et d'information visuelle caractérisant une image. Le système permet la recherche d'images ou de segments d'images par la communication au système de mots ou de caractéristiques d'images.

Une autre approche pour incorporer l'information apportée par un utilisateur au système est de mettre à jour interactivement une mesure de similarité significative de la sémantique d'un utilisateur *(Heinrichs et al., 2000)*. La fonction de similarité est une fonction pondérant les différentes caractéristiques. Les caractéristiques sont constituées entre autres par des points d'intérêt pris à différentes résolutions. Des similarités dites préattentives *(Santini et Jain, 1997)* sont mesurées entre les caractéristiques. Cette approche autorise plus de souplesse quant à la définition de la sémantique d'un utilisateur. L'utilisateur communique au système des exemples positifs et négatifs d'images. Ces exemples influencent directement sur les poids de la mesure de similarité, définissant ainsi la sémantique de l'utilisateur.

Un système nommé PicHunter utilise une approche bayésienne pour calculer la probabilité qu'une image d'une collection soit la cible d'un utilisateur prenant en compte l'historique de la recherche *(Cox et al., 2000)*. Le coeur de l'approche bayésienne est basé sur la modélisation probabiliste des actions d'un utilisateur lors de la recherche. L'interaction Homme-Machine se fait au travers d'une interface basique par laquelle l'utilisateur désigne des images similaires à sa cible. Suite à la communication au système de l'intérêt de l'utilisateur pour certaines images, le système met à jour les probabilités de toutes les images puis sélectionne un nombre donné d'images retournées à l'utilisateur. Deux schémas sont proposés pour effectuer cette sélection. L'un sélectionne les images associées aux probabilités les plus hautes, l'autre se base sur la maximisation de l'information via un critère entropique pour effectuer cette sélection.
Cette stratégie permet d'adapter la recherche à l'utilisateur. Celui-ci peut effectuer aussi bien la recherche d'une image cible spécifique, que la recherche d'images de même catégories ou même supporter une recherche ouverte comme celle effectuée par un utilisateur cherchant à acquérir de l'information sans nécessairement savoir exactement quelle information il cherche.

Un prototype nommé I2M a été développé par *Datcu et al. (1998)* et *Schroder et al. (2000)*. Ce système de fouille d'information dans les images est conduit par les connaissances d'un utilisateur qui communique son intérêt pour des structures de l'image spatialement localisées au niveau pixellaire ou au niveau de groupe de pixels. Cette nouvelle perspective autorise la recherche par le contenu, non pas seulement d'images entières, mais aussi de structures présentes dans la base de données ; la découverte d'information et la compréhension détaillée de la scène sont dès lors accessibles. Le système utilise une représentation hiérarchique de l'information formulée dans un contexte bayésien. Ce système a démontré son efficacité dans la fouille d'images multicapteur et multirésolution relatives à des domaines variés tels que l'imagerie satellitaire ou médicale. L'architecture d'un système opérationnel basé sur ce prototype est détaillée dans l'annexe A.1.
Dans une partie non-interactive sont effectuées l'extraction de caractéristiques (couleur, texture, etc) à partir des images et leur compression par classification non-supervisée. Des cartes spatiales de classification sont générées puis indexées dans une base de données relationnelle. De plus, des méta-informations textuelles de l'image sont insérées dans la base de données. Le clustering de l'espace des caractéristiques achève la modélisation non-supervisée des images sources.
Dans une partie interactive, un modèle sémantique est défini par des exemples positifs

et négatifs de type de structures spatiales par l'intermédiaire d'une interface graphique. Basée sur la représentation objective (en fonction des modèles employés) issue de la modélisation non-supervisée, la modélisation de la sémantique d'un utilisateur peut être effectuée en inférant les probabilités d'un réseau bayésien. Par ailleurs, la sémantique d'un utilisateur peut être sauvée et ainsi, des catalogues sémantiques peuvent être générés. Toujours interactivement, l'utilisateur peut effectuer une recherche probabiliste d'images à la sémantique similaire dans la base de données. Cette recherche est faite soit par la maximisation des probabilités a posteriori de la sémantique conditionnée par chaque image, soit par maximisation de la surface spatiale où la sémantique est présente ou soit par la maximisation d'un critère de séparabilité, significatif de la variance des probabilités a posteriori de la sémantique relatives aux pixels dans chaque image. Par ailleurs, une mesure entropique de pertinence est employée pour guider et évaluer la qualité de l'apprentissage.

La majorité des systèmes extraient des informations de couleur et de texture pour caractériser les images. Cette extraction d'information s'effectue au niveau pixellaire ou au niveau de points d'intérêt. La répartition spatiale des caractéristiques dans les images n'est pas assez souvent considérée. Or, elle permet une description plus approfondie du contenu des images. Toutefois, quelques approches, basées sur des modèles multirésolution ou sur des modèles hiérarchiques ou associatifs de clusters, décrivent l'information spatiale des images.
La modélisation sémantique est souvent issue de l'association de texte et d'information visuelle caractérisant une image. D'autres approches se basent sur la mise à jour d'une fonction de similarité, représentative de la conjecture de l'utilisateur ou sur l'apprentissage de relations hiérarchiques ou associatives de clusters.
L'information sémantique est incorporée dans le système par l'intermédiaire d'un langage naturel à l'utilisateur (exemples, degré de satisfaction, ...). Un apprentissage bayésien est souvent employé pour cette tâche.
La recherche probabiliste d'images de sémantique similaire s'effectue par des critères entropiques, de probabilité maximum, de séparabilité maximum ou encore de surface maximum. Il existe aussi des mesures entropiques de pertinence pour guider et évaluer la qualité de l'apprentissage.

3.3.3 La fouille d'information dans les séries temporelles

Extraction d'information dans les séries temporelles Un problème important intervenant dans la fouille d'information dans les séries temporelles est la définition de mesures de similarité pertinentes entre séries temporelles. Des modèles de similarité prennent en compte les distorsions entre séries temporelles, tels que le Dynamic Time Warping *(Berndt et Clifford, 1996)* ou encore l'appariement probabiliste de motifs *(Keogh et Smyth, 1997)*.
D'autres approches visent à évaluer des distances dans des espaces transformés (espace fréquentiel, ...) ou à évaluer des distances entre caractéristiques issues d'une modélisation stochastique de processus temporels. Sont inclus dans ces modèles stochastiques : les

chaînes de Markov cachées et les modèles de mélanges pour la modélisation linéaire de processus à état discret, les modèles de filtre de Kalman pour la modélisation linéaire de processus continus *(Kalman, 1960)*, ou encore le filtrage particulaire pour la modélisation de processus non-linéaire *(Gordon, 1993)*. Pour être significatives, ces méthodes d'analyse probabiliste nécessitent un nombre suffisamment grand d'échantillons. Par ailleurs, généralement ces modèles se basent sur un échantillonnage régulier. Les modèles de similarité entre motifs constituent des solutions plus directes pour l'extraction d'information de processus temporels par comparaison.

Modélisation sémantique des séries temporelles Plusieurs méthodes développées pour la modélisation sémantique découlent de l'extension aux séries temporelles du principe de la fouille de règles d'associations *(Agrawal, 1993)*. Mais, les modélisations sémantiques des séries temporelles sont majoritairement des classifications supervisées qui se basent sur des caractéristiques issues d'une modélisation stochastique du processus temporel. Or comme nous venons de le voir, une telle modélisation est difficile et en conséquence les classifications ne sont pas directement accessibles. Une alternative est le calcul de la similarité des séries temporelles par des exemples fournis par un utilisateur possédant une sémantique bien définie. Une méthode de fouille d'information dans les séries temporelles basée sur l'élaboration interactive d'un modèle de séquence par la fusion de façon itérative d'exemples positifs et négatifs de classes fournis par un utilisateur a été développée par *Keogh et Pazzani (1998)*. Elle emploie un facteur d'influence qui contrôle l'opérateur de fusion de séquences : un facteur d'influence positif autorise la généralisation du modèle et un facteur négatif entraîne l'exagération de la différence entre la séquence positive et négative.

3.3.4 La fouille d'information dans les séquences vidéo

La vidéo numérique est en plein essor pour la création, l'échange et le stockage d'information. Des méthodes efficaces pour l'indexation et la recherche d'information visuelle sont nécessaires pour permettre l'accès à de très grandes bases de données. Les vidéos sont généralement des signaux continus, présentant une scène composée de structures spatiales aux propriétés intrinsèques (forme, mouvement, texture, couleur ...). Le codage vidéo basé objet de la norme MPEG-4 utilise cette propriété générique pour coder et manipuler séparément des objets physiques dans un clip vidéo. La norme MPEG-7 apporte un niveau supérieur d'abstraction de l'information. Cette norme permet la création d'une description du contenu de la vidéo, comprenant selon la volonté de l'utilisateur, un niveau de détail variable. Cette description pourra contenir une information sur la syntaxe et la sémantique propre aux caractéristiques extraites des données ou propre aux relations entre les objets de la scène. Ces codages vidéo se prêtent bien aux techniques de recherche par le contenu.

Extraction d'information dans les vidéos Pour analyser et comprendre leurs contenus, la vidéo a besoin d'être segmentée en morceaux adaptés à une analyse perceptuelle. La plupart des systèmes actuels commencent par effectuer une segmentation temporelle de

la vidéo par un modèle hiérarchique de trames (images de base), sections (succession d'images décrivant un événement ou le comportement d'une structure) et de scènes (unité complète de narration).
Les transitions entre les sections peuvent être abruptes ou graduelles. La majorité des systèmes existants détectent les frontières entre les sections par des techniques de détection de changements entre trames ou paquets de trames successifs *(Borecsky et Rowe, 1996)*. L'analyse en composantes principales (ACP) a été employée par *Liu et Chen (2002)* pour l'extraction, à partir des précédentes trames, d'un vecteur de caractéristiques représentant le vecteur de la moyenne d'une distribution gaussienne. Un changement de scène est modélisé comme une différence de moyenne excédant un certain seuil entre deux trames consécutives.
Erol et Kossentini, (2000) exploitent le codage vidéo basé "objet" de MPEG-4 pour la sélection de trames clés. Cette sélection utilise l'information de mouvement des objets établie par les changements dans le mode de codage des textures et des formes.
La segmentation temporelle est suivie par la représentation et la modélisation du contenu de chaque section afin de permettre l'identification de sections similaires.
La plupart des systèmes n'utilisent qu'une seule trame de la section, appelée trame-clé, pour en représenter son contenu. Cette trame est souvent choisie arbitrairement. Pour les systèmes plus complexes qui sélectionnent plus d'une trame, des critères de changements multiples sont utilisés par *Zhang et al. (1997)* pour décider du nombre de trames clés. Pour la représentation et la mesure de similarité entre contenus de sections, l'utilisation de l'extraction et la représentation de sous-sections présentées par *Lin et al. (2001)*.
Suite à une phase d'extraction de caractéristiques à partir des trames-clés, des mesures de similarité sont définies pour prendre en compte la réelle similarité perceptuelle entre sections. Des distances euclidiennes définies dans L_2, des mesures de distance entre arbres ou encore des distances pseudo-sémantiques définies dans L_1 peuvent être utilisées *(Taskiran et al., 2000)*.
Les vidéos présentent beaucoup de similarité avec les STIS et beaucoup des méthodes d'extraction d'information dans la vidéo pourraient être adaptées aux STIS. Cependant, les STIS ont un contenu informatif qui diffère généralement beaucoup du contenu des vidéos : les échelles spatiales et temporelles de l'information utile ne sont pas les mêmes ; la quantité d'information est généralement beaucoup plus importante dans les STIS ; il existe un échantillonnage régulier dans un cas et irrégulier dans l'autre, etc.

Représentation sémantique des vidéos Les techniques développées essayent de combler le fossé sémantique entre l'information extraite des vidéos et l'interprétation qu'un utilisateur peut en faire.
Par une approche multimodale probabiliste (associant différentes sources d'information à la vidéo telles que du texte ou des bandes audio), un étiquetage sémantique dans un contexte bayésien a été développé par *Naphade et al. (1998)*. Cette modélisation sémantique ressemble à celle décrite dans le système *I2M* (annexe A.1). Etant donné le vecteur multimodal de la j-ème trame et prenant un a priori uniforme pour la présence/absence d'un concept A_ν, la probabilité d'occurrence de chaque concept A_ν dans la j-ème trame est obtenue par la loi de Bayes. Les probabilités conditionnelles $p(X_j \mid A_\nu)$ et $p(X_j \mid \neg A_\nu)$, avec

$\neg A_\nu$ représentant la sémantique négative[1], doivent être estimées, via une base d'apprentissage. Des labels sémantiques sont attribués aux concepts identifiés. Dans le même cadre, un apprentissage par région et la modélisation des interactions entre concepts sont décrits par *Naphade et al. (2002)*. Un réseau bayésien est utilisé pour modéliser les interactions entre les concepts. Cette approche permet de prendre en compte le contexte d'apparition.
Des relations temporelles de cause à effet ont été modélisées par *Tran et al. (2000)*. Une approche objet développée par *Khokhar et al. (1999)*, permet la modélisation inter et intra objets et ainsi, la définition de concepts.
Une grande similarité entre la vidéo et les STIS implique que certaines méthodes de représentation sémantique de vidéos pourraient être adaptées aux STIS.

3.3.5 La fouille visuelle d'information

Dans la fouille d'information, l'intégration d'un humain dans la boucle a de nombreux avantages (c.f. § 3.2.2). Dans cette optique, des techniques permettant des visualisations pertinentes de l'information pour afficher de grands volumes de données sont d'une grande importance.
La fouille visuelle d'information est le processus d'exploration et de découverte de connaissances, guidé par des représentations visuelles pertinentes.
Bien que l'approche traditionnelle utilise des visualisations seulement au début et à la fin de la recherche, cette stratégie ne tire pas avantage de l'information accessible et des interactions Homme-Machine possibles au travers de visualisations à l'intérieur même du processus de fouille. La visualisation peut constituer une étape à part entière dans le processus de fouille d'information. Par exemple, dans le système *I2M* (annexe A.1), l'apprentissage est conduit par l'interaction d'un utilisateur avec le système via des représentations dans l'espace image.
Cependant, les visualisations de l'information lors d'une fouille interactive peuvent être autres que des représentations dans l'espace des données. Des représentations visuelles du flux d'information transitant tout le long du canal de communication entre l'utilisateur et les données (c.f. § 3.2.3), constituent une information potentielle non utilisée, pouvant guider d'une façon parfois plus pertinente, l'utilisateur dans son exploration. En particulier, les espaces contenant l'information extraite par les modèles mathématiques, ou encore les modélisations sémantiques sont susceptibles de contenir une information visuelle pertinente. Par exemple, la visualisation de l'espace des caractéristiques d'une image peut aider à l'identification de structures particulières ayant une projection particulière dans l'espace image. Puis, la visualisation des clusters dans l'espace des caractéristiques, obtenus lors d'une classification permet une représentation plus intuitive des différents attributs de ceux-ci (moyennes, variances, densités, ...). Enfin, la visualisation des espaces probabilistes représentatifs du stade de l'apprentissage peut guider la fouille et aider à l'évaluation de sa qualité.

[1] Les exemples de sémantique négative désignent les contre-exemples d'un utilisateur.

Ces dernières années, beaucoup de techniques nouvelles de vision par ordinateur ont été développées et des techniques existantes ont été étendues pour s'adapter aux volumineuses bases de données et aux données multidimensionnelles. Dans la suite, une classification et une description des techniques importantes pour la fouille sont présentées.

Techniques orientées pixels L'idée de base des techniques orientées pixels est de cartographier les données dans des espaces images séparés. Chaque attribut de chaque donnée multidimensionnelle sera représenté par un pixel coloré dans l'image attribut correspondante. Il y aura donc autant d'images que d'attributs. Par exemple, en se référant toujours au concept *I2M*, les espaces images visualisées pourront être l'image des signatures spectrales et l'image des amplitudes radar.
Il existe deux types de techniques orientées pixels : les techniques de visualisation dépendantes et indépendantes de la recherche.
Les techniques dépendantes de la recherche arrangent les données en fonction de la recherche et les visualisations sont donc dynamiques. Pour ce faire, des distances sont calculées entre les données et les valeurs de la recherche, puis combinées pour chaque donnée permettant ainsi le calcul d'un critère général de pertinence. Les valeurs de pertinence relatives à chaques données sont ensuite arrangées d'une certaine façon dans les espaces images *(Keim et Kreigel, 1996)*. Il existe de nombreuses façons d'arranger des données ordonnées dans un espace image *(Morton, 1966)*. Généralement, les données pertinentes sont placées au centre de l'image et les données de pertinence décroissante sont arrangées en forme de spirale du centre vers l'extérieur de l'image. Toutefois, ces techniques, bien qu'adaptées pour la fouille de données ne possédant pas d'arrangement naturel, ne présentent pas beaucoup d'intérêt pour la visualisation de données possédant un ordre intrinsèque telles que les données images ou les séries temporelles. C'est donc pour cela que ces données seront visualisées par des techniques orientées pixels indépendants de la recherche qui arrangent les données dans l'espace image selon l'ordre intrinsèque des données (i.e. l'axe temporel, les axes spatiaux).

Techniques basées sur des projections géométriques Les techniques basées sur des projections géométriques visent à trouver des projections des données multidimensionnelles "intéressantes". Le critère évaluant l'intérêt peut être calculé par des propriétés statistiques des données telles que la corrélation pour l'analyse en composantes principales ou l'analyse factorielle, ou l'indépendance pour l'analyse en composantes indépendantes. Ces critères seront plus ou moins bien adaptés pour l'analyse selon la nature des données. Pour des STIS, ces techniques d'analyses sont utiles, parallèlement à la visualisation, pour l'analyse de caractéristiques multidimensionnelles. Pour un descriptif de ces techniques, on se réfèrera à la section sur la réduction de dimensionnalité (c.f. § 6.1.2).
Une autre technique basée sur des projections géométriques est la technique de visualisation par coordonnées parallèles *(Inselberg, 1990)*. Cette technique cartographie l'espace de dimension m dans l'espace bidimensionnel par l'emploi de m axes parallèles, chaque axe correspondant à une dimension. Les axes sont ordonnés selon un ordre artificiel ou inhérent aux données. Chaque donnée est représentée par une ligne polygonale qui intersecte les axes à une hauteur correspondante à la valeur de la donnée en cet attribut. C'est un principe simple mais qui est toutefois puissant pour révéler des caractéristiques

telles que la distribution des données et leurs dépendances. Cette technique est adaptée à la visualisation de multiples séries temporelles. Chaque coordonnée temporelle est représentée par un axe vertical distant du précédant proportionnellement à leur écart temporel. Du fait de la superposition de certains échantillons, cette représentation semble limitée pour un volume de données trop élevé. Une colorisation des lignes polygonales en fonction de la densité de données qu'elles supportent permettra cependant de visualiser des bases de données plus volumineuses.

Techniques basées sur des graphes Des techniques basées sur la visualisation des données sous formes de graphes permettent de représenter les dépendances entre les données ou structures.
Des systèmes ont été développés pour visualiser des structures en réseaux *(Becker, 1995)* ou pour rechercher des visualisations de données représentées sous formes de graphes *(Consens, 1993)*. Les techniques basées sur des graphes sont donc particulièrement bien adaptées à la visualisation des dépendances spatiales, des dépendances temporelles et des dépendances de caractéristiques.
La visualisation de la structure d'algorithmes modélisant les dépendances des données telles que les réseaux bayésien ou les réseaux de neurones peut permettre à l'utilisateur de comprendre les boîtes noires que sont souvent les algorithmes d'apprentissages. Dans le concept *I2M*, à défaut de visualiser l'ensemble des probabilités associées au réseau bayésien, des mesures évaluant la qualité de l'apprentissage sont présentées à l'utilisateur sous forme de classes d'histogrammes. Cette information visuelle renseigne l'utilisateur sur la contribution des différentes composantes dans l'apprentissage de façon à le guider dans sa recherche. Ce type de visualisation sera utile dans l'apprentissage sémantique des STIS.

3.4 Etat de l'art de l'analyse des séries temporelles d'images satellites

Aujourd'hui, les analyses de STIS sont limitées à des applications spécifiques. Les principales analyses sont la détection de changements et l'assimilation de données satellitaires pour l'extraction d'informations agronomiques de cultures, océaniques, etc.

3.4.1 Assimilation de données satellitaires

Les analyses des séries de données d'observation de la Terre étudient des évolutions complexes comme par exemple des évolutions agricoles, océaniques ou météorologiques.
Des modèles physiques permettent de décrire le fonctionnement de cultures, de l'océan, de l'atmosphère, etc, à partir de la connaissance de nombreuses caractéristiques propres

au domaine telles que la composition du sol et la biologie végétale pour l'agronomie ou la mécanique des fluides pour la météorologie. Malheureusement, certaines de ces caractéristiques sont mal connues et peuvent varier de manière significative dans l'espace et dans le temps. Ces modèles nécessitent donc l'apport d'autres informations. Les données satellitaires permettent de vérifier l'état réel des cultures, de l'océan, de l'atmosphère en tout point de l'espace, et de le comparer aux simulations des modèles de fonctionnement. En cas de discordance entre état réel et simulations, des corrections sont apportées aux paramètres mal connus du modèle. C'est le principe de l'assimilation des données satellitaires. Quelques exemples de méthodologies employées dans ce domaine sont présentés.

Modélisation du phénomène "El Niño" Beaucoup d'intérêt a été porté ces 15 dernières années sur l'étude de la STIS acquise par les satellites TOPEX et Poséidon. Elle présente des données altimétriques et de température du Pacifique équatorial Ouest. Cette STIS de basse résolution possède une haute fréquence d'échantillonnage temporel. L'observation des données brutes (altimétrie, température) au niveau de chaque pixel de la STIS a permis la validation de modèles physiques du phénomène issu du couple océan-atmosphère dénommé "El Niño", permettant ainsi la prédiction des dépressions tropicales associées. Les investigations de la NASA sur ce sujet sont disponibles sur Internet [84].
La basse résolution, qui autorise une haute fréquence d'échantillonnage temporel, permet l'analyse des données brutes, au niveau du pixel. De ces analyses découlent des modélisations phénoménologiques au niveau pixellaire utiles à la prévision. Mais l'avènement de la haute résolution, avec pour corollaire la constitution de STIS présentant des structures complexes, variées et irrégulièrement sous échantillonnés, ne permet pas les mêmes analyses car entre autre, il est difficile si ce n'est impossible d'adapter un modèle physique spécifique à chacune de ces nombreuses structures.

Assimilation de Données par Agro Modélisation (ADAM) Les STIS de haute résolution sont des sources d'information au potentiel énorme encore sous-exploitées. Etant donné la multitude de structures hétérogènes présentes dans de telles scènes, l'assimilation utilisant cette nouvelle génération de STIS se focalise sur la modélisation de certaines zones précises de la scène.
L'objectif du projet ADAM est de développer et évaluer une méthode permettant d'extraire des informations agronomiques sur les cultures (rendement potentiel, risques de maladie, besoins en eau ou en azote...) nécessaires pour la conduite optimale des cultures, pour estimer les productions et évaluer la qualité de l'environnement.
Pour mener à bien ce projet, un vaste dispositif expérimental a été mis en place pendant la campagne agricole 2000-2001 : acquisition massive d'images satellites optiques et radars, campagne de mesures terrain très précise (développement, état physiologique, rendement...), caractérisation fine du milieu (profils des sols, bilans hydriques et azotés ...). On se référera à la section 2.1 pour le détail de l'acquisition et des prétraitements effectués sur les images optiques SPOT contenues dans la base de donnée ADAM. Le projet se déroule dans la plaine du Danube en Roumanie, sur le site de Fundulea et est centré sur la culture du blé.
Une méthode a été développée pour l'estimation de l'état de la voûte de feuillage dans le temps et l'espace par l'utilisation de l'assimilation de données SPOT de haute fréquence

temporelle dans les modèles du processus de végétation *(Lauvernet et al., 2003)*.

L'approche n'est pas la même pour la fouille, qui se place dans un cas plus généraliste, et adopte une recherche exploratoire d'information, dédiée à presque n'importe quelle application. Cette différence majeure implique une grande divergence entre les techniques employées.

3.4.2 Détection de changements

La détection de changements est le processus d'identification d'états distincts d'une zone en l'observant à des dates différentes. Les analyses en détection de changements sont dédiées à la discrimination de deux classes de zones : "changements liés à un phénomène d'intérêt" et "autres". De plus, elles se limitent en général sur des données provenant de deux dates particulières.
La détection de changements peut se décomposer en deux étapes distinctes : l'obtention d'indices de changements suivie par la discrimination des zones de changements.

Obtention d'indices de changements Dans la première étape, on peut distinguer trois grandes classes d'indices employés dans la détection de changements :
Il existe tout d'abord des caractéristiques de changements ponctuels. Extraire des caractéristiques de différence d'images consiste tout simplement à calculer la différence pixel à pixel entre deux dates. Les zones sans changements apparaîtront avec des valeurs proches de zéro, alors que les zones soumises à des variations auront des valeurs absolues élevées. Un vecteur composé par des radiométries d'un même pixel à différentes dates engendre un espace de caractéristiques. L'analyse en composantes principales (ACP) de cet espace permet l'obtention de projections linéaires intéressantes pour la détection de changements. L'ACP s'interprètera différemment en fonction de la composition du vecteur engendrant l'espace de caractéristiques. L'espace des caractéristiques est composé des valeurs des pixels issues de :

- *Deux dates, un canal*. Si l'on suppose qu'il y a une relation linéaire entre les pixels qui n'ont pas changé entre les deux dates, on peut s'attendre à ce que ceux-ci résident sur une droite. Les changements apparaîtront alors dans la deuxième composante. Le problème est que, en ayant estimé d'abord la composante des pixels qui n'ont pas changé (la première composante), l'axe de la deuxième composante principale n'est pas forcément aligné avec les changements. De la même façon l'axe principal sera dévié par l'inertie des pixels qui ont changé *(Wang, 1993)*.
- *Plusieurs dates, un canal*. Pour certains chercheurs, l'information relative aux changements dans une STIS est considérée comme un bruit et elle doit apparaître dans les dernières composantes principales. Evidemment, cette approche n'est pas robuste, car la notion de bruit dépendra du pourcentage de pixels ayant changé. Pour éviter la confusion avec le bruit, la différence d'images est considérée en entrée de l'ACP. Etant donné que les images différences contiennent des changements, ceux-ci auront tendance à se concentrer sur la première composante. De la même façon, on peut employer en entrée de l'ACP n'importe quelle mesure de changements analogue à la

différence d'images entre des dates consécutives.
- *deux dates, plusieurs canaux*. Afin de concentrer l'information relative aux changements d'un couple multispectral sur une seule bande, l'ACP est employée avec les n différences bande à bande. Fung propose de faire une ACP avec l'ensemble des $2n$ images *(Fung, 1990)*. Le problème est que les bandes sont traitées de la même façon indépendamment de leur date.

Le principal désavantage de ces méthodes, est qu'elles sont très sensibles au bruit et imposent donc, a priori, un étalonnage des données.
En second lieu, il existe des caractéristiques de changements dans un voisinage pixellaire. Afin de prendre en compte les changements des interactions spatiales des pixels, des méthodes de caractérisation des textures sont employées. Elles mesurent les dépendances entre les pixels et leurs voisinages. Les méthodes basées sur l'ACP décrites précédemment peuvent être ensuite employées de la même façon avec les caractéristiques de texture. La corrélation spatiale entre deux dates ou encore l'information mutuelle peut être employée pour quantifier les changements à l'intérieur d'une fenêtre.
Enfin, il existe des caractéristiques de changements à l'échelle des structures. Une approche proposée par *Bruzzone et Prieto (2000a)*, consiste à segmenter les deux images (avant et après) et fusionner les segmentations. Cette fusion consiste à obtenir une segmentation commune aux deux images où chaque parcelle est homogène. Ensuite, chaque parcelle est caractérisée par un vecteur d'indices de changements qui permet d'évaluer les changements en restant à la résolution des parcelles. Cette méthode engendre des résultats détectant des zones de changements compactes et a pour avantage d'être robuste au bruit. Dans ce même type d'approche, nous pouvons citer l'utilisation de la classification d'images. En effet, les changements peuvent être identifiés comme étant les pixels qui changent de classes d'une image à l'autre. Par cette approche, on peut s'affranchir des problèmes difficiles d'étalonnage et de recalage difficile des images. Toutefois, il faut être capable d'extraire les structures d'intérêts ce qui nécessite généralement des connaissances a priori.

Inférence d'une classe de changement Une fois que le degré de changements a été évalué, l'on est amené à décider si la zone considérée a changé ou pas. Des méthodes de classifications automatiques ou supervisées doivent déterminer par le choix d'un seuil approprié si, étant donnée une valeur de changement, celui-ci a effectivement eu lieu. L'approche bayésienne est l'approche majoritairement utilisée pour fixer ce seuil.
On peut poser le problème de détection de changements comme un problème de classification. La règle de Bayes permet de choisir le seuil optimal qui sépare deux classes en connaissant leurs distributions de probabilités. La difficulté réside alors à trouver ces lois de probabilité. Deux possibilités existent :
- l'algorithme apprend ces lois à l'aide de régions d'apprentissage pour lesquelles on connaît la classe. C'est le cas supervisé.
- l'algorithme estime ces lois à partir des données sans connaître la classe d'aucun pixel. C'est le cas non supervisé.

Un exemple où l'on accepte l'intervention d'un utilisateur (classification supervisée) consiste à faire une classification conjointe des deux images en fixant le nombre de classes égal à 2 et à faire un apprentissage sur des zones pour lesquelles l'on connaît la vérité

terrain.
Un autre exemple est proposé par *Smits et Anoni (2000)*. Des zones de changements sont détectées par la communication au système de la probabilité de non-détection et de la probabilité de fausse alarme.
Dans le cas automatique (classification non-supervisée), la façon la plus simple de procéder est de faire une classification sur les deux dates indépendamment, puis de calculer la différence des classifications. Cette approche dépend beaucoup des performances des algorithmes de classification utilisés.
Bruzzone et Prieto (2000b) proposent deux approches basées sur la règle de Bayes. Ils partent d'une image différence entre deux dates pour laquelle ils supposent avoir deux classes de pixels : les pixels où il y a des changements ω_c et ceux où il n'y en a pas ω_n. La probabilité de la valeur X de chaque pixel de l'image de différence peut être écrite comme : $p(X) = p(X \mid \omega_n)p(\omega_n) + p(X \mid \omega_c)p(\omega_c)$. Sous cette hypothèse, l'algorithme EM (c.f. § 4.5) est utilisé afin d'estimer $p(X \mid \omega_n)$, $p(\omega_n)$, $p(X \mid \omega_c)$ et $p(\omega_c)$. Ils supposent que les deux probabilités conditionnées suivent des lois gaussiennes, ce qui simplifie l'algorithme à l'estimation de la moyenne et de la variance des lois.
Ensuite, les dépendances spatiales entre pixels sont prises en compte (cas adapté aux capteurs à haute résolution). Les champs de Markov (c.f. § 6.1.1.1) sont utilisés pour modéliser les dépendances entre pixels par définition de fonctions de potentiel. Des probabilités de dépendance sont ainsi introduites dans l'algorithme EM pour l'estimation des probabilités de changement et non-changement.
L'approche est robuste au bruit et détecte des zones compactes.
La méthode est étendue au cas multicapteur et au cas des capteurs pour lesquels la distribution des différences n'est pas forcément gaussienne par l'utilisation d'un modèle paramétrique.

L'extraction de caractéristiques dans la détection de changements se focalise sur l'obtention d'indices significatifs de changements ayant lieu principalement entre deux dates. Pour la fouille, les intérêts sont plus vastes. Ils se portent sur des méthodes d'analyses exploratoires incluant la détection de changements, mais englobant aussi l'identification d'autres types d'évolutions caractérisées par exemple par des variations plus lentes sur une série d'échantillons temporels successifs. En conséquence, dans notre cas, l'utilisation par exemple de la seule différence d'images ou des dernières composantes d'une ACP est trop restrictive pour caractériser tous les types d'événements possibles. Toutefois, les méthodes de quantification de changements développées constituent des caractérisations du signal spatio-temporel intéressantes pour la fouille. En particulier, l'analyse de changement à l'échelle des structures issues d'une segmentation ou de variation de texture est intéressante.
Par ailleurs, la détection de changements se concentre sur la discrimination de deux classes d'événements. Celle-ci pourra être étendue à la discrimination de plus de deux classes de signaux pour être utile à la fouille.
Enfin, dans la littérature, la domination des méthodes bayésiennes pour effectuer cette discrimination atteste de la puissance de l'approche.

3.5 Fouiller l’information dans les séries temporelles d’images satellites

Ces dernières décennies, de grandes quantités d’images de beaucoup de sites terrestres ont été acquis par les satellites, permettant ainsi la constitution de séries temporelles d’images satellites. De plus, les opportunités pour générer de nouvelles séries sont grandissantes : les satellites, de plus en plus nombreux et avec des vitesses d’acquisition de plus en plus rapide, permettent l’augmentation de la fréquence d’acquisition d’une même scène. Ainsi, l’observation de la dynamique des scènes et des structures spatio-temporelles est de plus en plus accessible.

Mais, comme pour les images satellites, les méthodes traditionnelles d’interprétation de séries temporelles d’images emploient à l’heure actuelle des experts qui inspectent visuellement les données pour une application particulière. Ceci limite la pleine exploitation des archives.

Des approches exploratoires sont nécessaires pour dégager des connaissances dans les nombreux domaines d’application. Pour répondre à cette problématique, le paradigme de la fouille peut être appliqué aux séries temporelles d’images satellites. Jusqu’à présent, ce champ d’investigation est encore quasi-inexploré. Le processus de fouille d’information dans une série temporelle d’images satellites implique la recherche de motifs spatio-temporels dans les données.
Des requêtes typiques de la fouille pourraient se formuler par "trouver des évolutions de couverture forestière identiques à celles-ci", ou "trouver toutes les moissons effectuées entre juin et septembre", ou bien "quelles sont les similarités et les différences d’évolutions entre ces deux structures spatio-temporelles ?" ou encore "quels sont les comportements typiques de rivières en hiver ? Et en été ?".

Mais, cette tâche n’est pas aisée. A partir de la caractérisation des données faite au chapitre 2.3, et de l’état de l’art de la fouille d’information et de l’analyse des séries temporelles d’images satellites, qui a été présenté dans les précédentes sections, nous formulons un cadre de travail pour le développement d’un concept de fouille d’information dans les séries temporelles d’images satellites.

Contraintes sur la modélisation des séries temporelles d’images satellites

Comme nous l’avons vu dans la section 3.2.3, la fouille d’information se compose de deux étapes distinctes dont la première consiste à extraire des informations de la source par modélisation du signal. Cette première phase de traitements a pour objectif de modéliser les structures spatio-temporelles par l’utilisation de librairies d’algorithmes afin d’engendrer une description la plus complète possible de la scène dynamique.

Basé sur les chapitres précédents, nous énumérons les contraintes sur la modélisation des séries temporelles d'images satellites imposées par l'approche de la fouille d'information.
Premièrement, du fait de l'hétérogénéité des motifs spatio-temporels observés (c.f. § 3.4), la modélisation doit davantage s'orienter vers une description du signal spatialement et temporellement localisée, plutôt que vers une description globale de la scène comme c'est le cas dans certaines recherches d'images ou de vidéos par le contenu.
Deuxièmement, la modélisation de signaux multidimensionnels est un problème difficile. De plus, il nous faudra modéliser des dépendances spatio-temporelles. Comme nous le verrons dans le chapitre suivant, l'approche bayésienne présente de nombreux avantages pour réduire la complexité de tels problèmes. L'état de l'art démontre que les méthodes bayésiennes sont déjà largement employées dans ces situations.
Ensuite, les caractéristiques de couleur, de texture et de forme sont les caractéristiques les plus largement employées dans la fouille d'images et de vidéos. Mais, contrairement à la fouille dans les séries temporelles, l'extraction directe de caractéristiques d'évolutions temporelles n'est pas facile : l'échantillonnage irrégulier et la faible quantité d'échantillons ne facilitent pas la modélisation stochastique des processus temporels.
La modélisation de la dynamique de la scène au niveau pixellaire est difficile (c.f. § 2.3). Pour cela, il faut prendre en compte l'échantillonnage irrégulier des motifs qui sont souvent sont sous-échantillonnés, ainsi que les discontinuités temporelles des radiométries et l'inexactitude de la superposabilité géométrique.
Finalement, une analyse par région sera plus robuste aux erreurs de superposabilité géométrique et utile pour l'analyse de motifs sous-échantillonnés temporellement. Les évolutions spatio-temporelles des différentes structures de la scène pourront être caractérisées par des méthodes à la confluence de la recherche de vidéos par le contenu et de la détection de changements.

Modélisations sémantiques des séries temporelles d'images satellites

Dans le paradigme de la fouille (c.f. § 3.2), la modélisation de la sémantique a pour objectif d'établir des liens entre les données modélisées et les intérêts des utilisateurs. Nous tirons quelques conclusions à partir de l'état de l'art dans le but de développer un modèle sémantique adapté aux séries temporelles d'images satellites.

Comme le montre l'état de l'art, les modèles bayésiens sont efficaces pour coder l'incertitude de connaissances expertes dans les systèmes de fouille d'information. De plus, une approche bayésienne permet de modéliser des sémantiques complexes, comme les dépendances spatio-temporelles ou des dépendances de caractéristiques.
Des sémantiques définis par association ou hiérarchie de clusters, ou encore par la définition interactive de similarité en fonction d'un motif de référence, ont déjà démontré leurs performances (c.f. § 3.3).
Par ailleurs, l'état de l'art sur la fouille d'information dans les vidéos ou sur la détection de changements a montré qu'une décomposition de la scène en objets pouvait permettre d'inclure leurs relations spatio-temporelles, dans le modèle sémantique.
Pour définir l'intérêt ou la conjecture d'un utilisateur, plusieurs systèmes présentés dans

l'état de l'art estiment interactivement les probabilités associées aux modèles sémantiques via la communication au système de quelques exemples positifs ou négatifs éventuellement nuancés.
Enfin, des mesures de pertinence entropiques sont couramment utilisées pour quantifier la taille du fossé sémantique.

Représentations visuelles pour la fouille dans les séries temporelles d'images satellites

Les séries temporelles d'images satellites constituent des signaux complexes multidimensionnels. L'information est disséminée ici au travers des différentes sources constituées par chaque pixel et en chaque temps. De ce fait, certaines informations présentes dans ce signal sont parfois difficilement accessibles par le système de perception visuelle humaine. En effet, la section 3.2.2 de ce chapitre a mis en évidence les difficultés humaines à mesurer l'information absolue et à analyser de grandes quantités d'information multidimensionnelle. Des représentations visuelles de plus faible dimensionnalité et rehaussant les dépendances temporelles, spatiales et entre caractéristiques, sont nécessaires pour une meilleure interprétation. De même, des représentations visuelles accompagnant le processus d'apprentissage semblent utiles pour maximiser l'interaction Homme-Machine.

Certaines représentations présentent un intérêt pour la visualisation. Premièrement, des représentations dans l'espace des caractéristiques évolutives $\mathcal{E}_{t,\theta}$ révèlent l'information sur l'évolution de la distribution des caractéristiques. Deuxièmement, pour représenter l'information complémentaire sur l'évolution spatiale des structures (c.f. § 2.4), on pourra aussi utiliser l'espace spatio-temporel $\mathcal{E}_{r,t}$. Ensuite, l'espace MT des caractéristiques est un espace de dimension élevée. Des projections de celui-ci seront utiles pour une analyse exploratoire visuelle (c.f. § 3.3.5). En particulier, afin de décomposer l'évolution des caractéristiques du signal dans le temps, on pourra employer des projections dans les espaces TL de caractéristiques $\mathcal{E}_{\theta_{t_i}}$. Par ailleurs, les caractéristiques radiométriques des STIS représentent des signatures spectrales dans les canaux vert, rouge et proche-infrarouge. Ce sont des mesures physiques, associées à un certain canal dans le visible. Les experts ont l'habitude d'employer pour l'interprétation, l'espace de couleur rouge-vert-bleu (RGB). Il n'est donc pas souhaitable de changer l'espace de couleur. Par contre, en peut employer le même espace de représentation $\mathcal{E}_{r,t,\theta}$, mais adapter la dynamique des couleurs des images pour réhausser l'homogénéité visuelle des radiométries entre les différentes images des séries temporelles. Enfin, comme précisé dans la section 3.3.5, des représentations visuelles d'information liées au processus d'apprentissage pourront guider le processus de fouille.

3.6 Résumé

Dans ce chapitre, les points suivants ont été discutés :

- après avoir exposé les principales motivations, le concept de la fouille d'information a été défini ; en utilisant une analogie avec le problème du canal de communication,

nous avons décomposé le processus de fouille d'information en deux parties : la modélisation du signal et la modélisation sémantique ;

- un état de l'art sur la fouille d'information et sur l'analyse des STIS a été dressé et discuté ;

- grâce à cet état de l'art et aux conclusions tirées du chapitre 2, nous avons défini les objectifs de la fouille d'information dans les STIS et établi un cadre de travail pour y parvenir.

Deuxième partie

Apprentissage bayésien de structures spatio-temporelles

Chapitre 4

Inférence bayésienne pour la modélisation de signaux multidimensionnels

La modélisation stochastique de signaux multidimensionnels permet l'extraction et la représentation de connaissances à partir de données et de connaissances expertes. Les méthodes d'inférence constituent le coeur de la modélisation. En effet, des processus d'inférence sont utilisés pour évaluer la qualité des modèles représentatifs, c'est à dire pour quantifier comment des modèles paramétriques s'ajustent aux données. En conséquence, pour parvenir à des représentations pertinentes de l'information contenue dans les séries temporelles d'images satellites et ainsi répondre à la problématique de la fouille, il est crucial d'inférer les "bons" modèles représentatifs. L'inférence bayésienne se propose de répondre à cette problématique.

L'inférence bayésienne a prouvé son efficacité pour l'extraction d'information à partir de signaux multidimensionnels. Cette efficacité résulte de plusieurs propriétés. Premièrement, les méthodes bayésiennes permettent le traitement de données incomplètes. Deuxièmement, la grande complexité des modèles construits sur un grand nombre de variables inter-dépendantes peut être abordée en considérant et en exploitant des hypothèses d'indépendance conditionnelle propres à un réseau bayésien. Ce dernier est représenté par un graphe dont les noeuds représentent les variables stochastiques et les arêtes codent les dépendances entre variables. Troisièmement, les connaissances a priori et les données peuvent être naturellement combinées dans les techniques de statistique bayésienne. Enfin, les méthodes bayésiennes proposent une approche efficace pour limiter la complexité d'une modélisation et sélectionner des modèles adéquats. Dans une première partie, le formalisme bayésien sera donc exposé.
Les méthodes entropiques, basées sur la quantification de l'information, permettent aussi la réalisation d'inférences dans une certaine mesure équivalentes avec l'approche bayésienne.

Les différentes mesures de l'information ainsi que les différentes méthodes entropiques pour l'extraction d'information seront exposées à la suite des méthodes bayésiennes. Les équivalences entre les méthodes entropiques et bayésiennes seront aussi clarifiées. Par ailleurs, comme il est important de définir des familles de modèles puissants sur lesquels s'appliquera l'inférence bayésienne, on présentera des modèles stochastiques pour la représentation de signaux multidimensionnels. Enfin, on abordera le problème de l'optimisation dans le processus de modélisation, en présentant des algorithmes d'optimisation déterministes et stochastiques.

4.1 Approche bayésienne des probabilités et des statistiques

Pour étudier l'approche bayésienne, quelques notations sont introduites. On note les variables aléatoires avec des majuscules (X, Y, X_i, Θ) et l'état ou la valeur d'une réalisation particulière avec des minuscules (x, y, x_i, θ). On note un jeu de variables par des majuscules en caractères gras ($\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{X_i}, \boldsymbol{\Theta}$) et un jeu de réalisations par les correspondantes minuscules ($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x_i}, \boldsymbol{\theta}$). On utilise $p(X = x \mid \xi)$ ou plus simplement $p(x \mid \xi)$ pour désigner la probabilité de l'état $X = x$ ou la fonction de distribution de probabilités de X conditionnée par les connaissances ξ. La signification de cette dernière notation sera facilement clarifiée par le contexte.
La définition bayésienne de probabilité d'un événement x est la représentation par des nombres réels de son degré de plausibilité relativement à un stade de connaissances. Cette définition est différente de la définition fréquentiste des probabilités dans le sens que la probabilité $p(x)$ n'est pas évaluée sur la seule base de tirages répétés mais est calculée considérant un conditionnement par des connaissances a priori ξ. La probabilité conditionnelle est définie par

$$p(x \mid \xi) = \frac{p(x, \xi)}{p(\xi)}.$$

Une conséquence immédiate de la définition de probabilités conditionnelles est la loi de Bayes. Pour deux variables aléatoires X et Y, elle s'écrit

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \tag{4.1}$$

Cette formule permet l'inversion des probabilités conditionnelles et en conséquence autorise le calcul de la probabilité a posteriori $p(x \mid y)$ de la variable X à partir de sa vraisemblance $p(y \mid x)$, de la probabilité a priori $p(x)$ et de la normalisation introduite par le terme $p(y)$. La loi peut être vue comme la mise à jour de la probabilité a priori de la variable X à partir de l'assimilation d'une nouvelle information sur la variable Y.

4.1.1 Réseaux bayésiens

Un réseau bayésien pour une collection de variables $\mathbf{X} = \{X_1, ..., X_n\}$ consiste en (1) une structure de graphe $\mathcal{G}$ dirigée et acyclique codant des affirmations sur l'indépendance

conditionnelle des variables $\mathbf{X}$, et en (2) une collection $\mathcal{P}$ de distributions de probabilités locales associées à chaque variable *(Heckerman et al. 2000)*. Les noeuds de $\mathcal{G}$ correspondent aux variables $\{X_i\}$ et les arcs de $\mathcal{G}$ codent les dépendances entre chaque variable X_i et la collection correspondante de variables dont celle-ci dépend. On appelle les variables dont dépend X_i ses variables parents pa_{X_i}.
La paire $(\mathcal{G}, \mathcal{P})$ code la distribution conjointe $p(\mathbf{x})$. Cette dernière est exprimée par le produit de distributions de probabilités locales :

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid pa_{x_i}). \tag{4.2}$$

En effet, par la règle de chaîne de probabilités, on a $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, ..., x_{i-1})$. Or si la variable X_i ne dépend que des ses parents pa_{x_i} qui représente un sous-ensemble de variables, on a la propriété de Markov

$$p(x_i \mid x_1, ..., x_{i-1}) = p(x_i \mid pa_{x_i}) \tag{4.3}$$

qui rapportée dans la règle de chaîne de probabilités, prouve l'Eq.4.2.
Parce qu'un réseau bayésien pour $\mathbf{X}$ détermine la distribution de probabilités conjointes de $\mathbf{X}$, on peut l'utiliser pour évaluer facilement n'importe quelle probabilité en exploitant les propriétés d'indépendance conditionnelle. L'évaluation de ces probabilités est un processus d'inférence.

L'apprentissage des probabilités d'un réseau bayésien à partir de données comprend une collection de techniques d'analyse de données combinant les connaissances a priori d'un utilisateur avec les données pour produire des connaissances améliorées. Considérant les vecteurs de paramètres $(\Theta_1, ..., \Theta_i, ...)$ intervenant dans la modélisation des distributions de probabilités conditionnelles $p(x_i \mid pa_{x_i}, \theta_i, \mathcal{G})$, et en utilisant l'Eq. 4.2 on peut calculer la distribution conjointe $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \mathcal{G})$. On code l'incertitude sur l'ensemble des vecteurs de paramètres $\boldsymbol{\theta}$ en définissant une variable aléatoire $\boldsymbol{\Theta}$ et une distribution de probabilités a priori $p(\boldsymbol{\theta} \mid \mathcal{G})$ associée. Le problème d'apprentissage se résume alors à l'inférence de la distribution de probabilités a posteriori $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{G})$. Le cas de l'apprentissage des probabilités d'un réseau bayésien utilisant une distribution multinomiale pour $p(x_i \mid pa_{x_i}, \theta_i, \mathcal{G})$ a été largement étudié, notamment par *Cooper et Herskovits (1992)*. Ce modèle sera détaillé dans le chapitre sur l'apprentissage interactif de la sémantique d'un utilisateur (c.f. § 7.1.2).

4.1.2 Modèles hiérarchiques bayésiens

Une hiérarchie de variables $X_0 \rightarrow X_1 \rightarrow ... \rightarrow X_n$, caractérisant des données X_0 à des niveaux sémantiques de plus en plus raffinés, constitue un modèle hiérarchique. Ce dernier est dit bayésien si la distribution conjointe vérifie la relation

$$p(x_0, ..., x_n) = p(x_n \mid x_{n-1})p(x_{n-1} \mid x_{n-2})...p(x_1 \mid x_0)p(x_0). \tag{4.4}$$

C'est un cas particulier d'un réseau bayésien (Eq. 4.2), où l'on considère que la distribution d'une variable X_i à un certain niveau i, dépend de la variable X_{i-1} du niveau inférieur et, conditionnellement à cette variable X_{i-1}, est indépendante de toutes les variables

$X_{i-2}, X_{i-3}, ..., X_0$ de niveau inférieur.
Par exemple, si l'on considère un modèle hiérarchique bayésien à trois niveaux $X_1 \rightarrow X_2 \rightarrow X_3$, la vraisemblance $p(X_1 \mid X_2, X_3) = p(X_1 \mid X_2)$, ce qui veut dire que si l'on a la connaissance de X_2, l'information apportée par X_3 n'intervient pas dans le calcul de la vraisemblance. Cette hypothèse est raisonnable car l'information sur X_3 est introduite dans le calcul de l'a priori $p(X_2)$ par l'intégrale de marginalisation

$$p(x_2) = \int p(x_2 \mid x_3)p(x_3)dx_3. \tag{4.5}$$

La règle de Bayes peut être employée pour le passage d'un niveau à l'autre de la hiérarchie en incorporant un a priori. Plus précisément, l'évaluation de la probabilité a posteriori $p(x_i \mid x_{i-1})$ d'une variable X_i utilise la vraisemblance $p(x_{i-1} \mid x_i)$ et la probabilité a priori $p(x_i)$.
Un modèle hiérarchique bayésien pourra formaliser le raffinement progressif de l'information impliqué dans le processus de fouille (c.f. § 5.1).

4.1.3 Deux niveaux de l'inférence bayésienne

Le calcul d'une probabilité d'intérêt étant donné un modèle est généralement nommé *inférence probabiliste*. Les techniques bayésiennes dans le traitement des images et des séries temporelles tirent leur puissance du fait qu'il existe de nombreux modèles paramétriques adaptés, sur lesquels peut s'appliquer une inférence probabiliste.
Considérons une collection finie de N modèles $\mathcal{M} = (\mathcal{M}_1, ..., \mathcal{M}_N)$, possédant chacun un vecteur de paramètres aléatoires Θ_i, $i \in [1, N]$. Considérant les données X, les vecteurs de paramètres Θ et les modèles $\mathcal{M}$, on peut construire un modèle hiérarchique bayésien à 3 niveaux

$$X \rightarrow \Theta \rightarrow \mathcal{M}$$

avec les probabilités conditionnelles $p(x \mid \theta_i, \mathcal{M}_i)$ et $p(\theta_i \mid \mathcal{M}_i)$ associées à chacun des N modèles.

Le *premier niveau d'inférence* est l'évaluation des probabilités du vecteur de paramètres étant donnée l'observation des données x et un modèle $\mathcal{M}_i$. Utilisant la loi de Bayes (Eq. 4.1), on peut estimer la probabilité a posteriori de Θ_i à partir de sa vraisemblance et de sa probabilité a priori :

$$p(\theta_i \mid x, \mathcal{M}_i) = \frac{p(x \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i)}{p(x \mid \mathcal{M}_i)} \propto p(x \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i) \tag{4.6}$$

La probabilité $p(x \mid \mathcal{M}_i)$ est une constante de normalisation ignorée dans ce premier niveau car elle n'influe pas sur l'inférence de Θ_i.

Le *deuxième niveau d'inférence* est le calcul de la probabilité a posteriori des modèles $\mathcal{M}_i$.

$$p(\mathcal{M}_i \mid x) = \frac{p(x \mid \mathcal{M}_i)p(\mathcal{M}_i)}{p(x)} \propto p(x \mid \mathcal{M}_i)p(\mathcal{M}_i) \tag{4.7}$$

avec l'a priori sur le modèle $p(\mathcal{M}_i)$ et l'a priori sur les données $p(x)$. L'évidence $p(x \mid \mathcal{M}_i)$, révélatrice de la vraisemblance d'un modèle $\mathcal{M}_i$ pour des données X, est un terme prépondérant dans le calcul de la probabilité a posteriori du modèle contrairement à l'Eq. 4.6 où il joue le rôle d'une constante de normalisation. Il peut être obtenu par marginalisation, c'est à dire par intégration de l'espace du vecteur de paramètres Θ_i.

$$\begin{aligned} p(x \mid \mathcal{M}_i) &= \int p(x, \theta_i \mid \mathcal{M}_i) d\theta_i \\ &= \int p(x \mid \theta_i, \mathcal{M}_i) p(\theta_i \mid \mathcal{M}_i) d\theta_i \end{aligned} \tag{4.8}$$

Notons que cette intégration ne peut se calculer analytiquement que pour certains modèles, et que son calcul numérique est souvent trop coûteux. Toutefois, nous verrons par la suite que des approximations permettent son évaluation.

Le choix des distributions de probabilités a priori influence directement les deux niveaux d'inférence. Il convient donc de choisir un a priori approprié à la quantité d'information disponible.

- Si aucune information n'est disponible, des a priori non-informatifs issus par exemple du principe de l'entropie maximum *(Jaynes, 2003)* sont employés (c.f. § 4.3.1). Par exemple, pour un a priori sur un modèle de mélange de gaussiennes, la distribution a priori de l'écart type σ de chaque gaussienne devra être la distribution a priori de Jeffreys $p(\sigma) \propto 1/\sigma$, et la distribution a priori de chaque moyenne μ devra être une gaussienne de variance suivant aussi une distribution a priori de Jeffreys. Toutefois, on se ramène le plus souvent au cas simple d'une distribution a priori uniforme (c.f. § 6.1.3).
- Lorsque l'inférence probabiliste est conduite par un apprentissage interactif (e. g. c.f. § 7.1.2), l'emploi d'un a priori conjugué est utile pour la mise à jour progressive d'un a posteriori. Un tel a priori, multiplié par la vraisemblance engendre un a posteriori conservant la forme de l'a priori. Ainsi, l'a posteriori engendré par un jeu de données peut être réutilisé comme a priori lors des apprentissages suivants.

4.1.4 Estimation de paramètres

L'inférence probabiliste s'attache à évaluer des distributions de probabilités relatives à une variable Θ à partir de la connaissance des distributions de probabilités de x. La théorie de la décision, et en particulier l'estimation de paramètres, se préoccupe quant à elle de choisir des valeurs scalaires souvent nécessaires à la prise de décision. En estimation de paramètres, cette valeur scalaire est notée $\widehat{\theta}$.

Risque de Bayes Dans l'estimation d'un paramètre aléatoire, une fonction de coût $c(\theta - \widehat{\theta})$ sert à mesurer l'importance de l'erreur d'estimation. Les principales fonctions de coût utilisées sont le coût quadratique

$$c_q(\theta - \widehat{\theta}) = (\theta - \widehat{\theta})^2 \tag{4.9}$$

et le coût uniforme

$$c_u(\theta - \widehat{\theta}) = \begin{cases} 0 & \text{si } |(\theta - \widehat{\theta})| < \frac{\delta}{2} \\ 1 & \text{si } |(\theta - \widehat{\theta})| > \frac{\delta}{2} \end{cases} \tag{4.10}$$

L'espérance du coût de la distribution conjointe $p(x, \theta)$ est appelée risque de Bayes

$$R = E[c(\theta - \widehat{\theta})] = \int \int c(\theta - \widehat{\theta}) p(x, \theta) dx d\theta. \tag{4.11}$$

La minimisation de ce risque conduit à des estimateurs bayésiens. Le minimum du coût quadratique correspond à l'estimateur du minimum de la moyenne de l'erreur quadratique et pour le minimum du coût uniforme correspond à l'estimateur du maximum a posteriori.

Estimateur du minimum de l'erreur quadratique moyenne (MMSE) Le minimum du risque associé au coût quadratique $c_q(\theta - \widehat{\theta})$

$$R_q = \int p(x) dx \int (\theta - \widehat{\theta})^2 p(\theta \mid x) d\theta$$

est atteinte en utilisant l'estimateur MMSE

$$\widehat{\theta}_{MMSE} = \int \theta p(\theta \mid x) d\theta. \tag{4.12}$$

L'estimateur MMSE n'est autre que la moyenne de la distribution de probabilités a posteriori $p(\theta \mid x)$. Cet estimateur sera donc robuste en présence de distributions bruitées.

Estimateur du maximum a posteriori (MAP) La minimisation du risque associé au coût uniforme $c_u(\theta - \widehat{\theta})$

$$R_q = \int p(x) dx \Big(1 - \int_{\widehat{\theta} - \frac{\delta}{2}}^{\widehat{\theta} - \frac{\delta}{2}} p(\theta \mid x) d\theta \Big)$$

est atteinte par l'estimateur MAP

$$\widehat{\theta}_{MAP} = arg \max_{\theta} p(\theta \mid x). \tag{4.13}$$

L'estimateur MAP est donc le maximum de la distribution a posteriori $p(\theta \mid x)$. En conséquence, si la distribution a posteriori est symétrique les estimateurs $\widehat{\theta}_{MMSE}$ et $\widehat{\theta}_{MAP}$ sont équivalents. Ces estimateurs se basent sur la distribution a posteriori. Ils sont donc dépendants de la distribution a priori $p(\theta)$.

Estimateur du maximum de vraisemblance (ML) Lorsque le paramètre inconnu à estimer est déterministe la distribution a priori $p(\theta)$ est un Dirac et le risque de Bayes ne peut plus être employé. On emploie alors la fonction de vraisemblance $p(x \mid y)$. Elle représente l'incertitude dans les mesures. L'estimateur du maximum de vraisemblance, défini par

$$\widehat{\theta}_{ML} = arg \max_{\theta} p(x \mid \theta), \tag{4.14}$$

est le plus pertinent pour l'estimation de paramètres déterministes ou lorsque, pour un paramètre aléatoire, aucun a priori n'est disponible. On remarque que l'estimateur $\widehat{\theta}_{ML}$ est égal à l'estimateur $\widehat{\theta}_{MAP}$ pour une distribution a priori $p(\theta)$ uniforme et à l'estimateur $\widehat{\theta}_{MMSE}$ pour une distribution a priori $p(\theta)$ symétrique. Bien que l'estimateur $\widehat{\theta}_{ML}$ ne soit pas basé sur le risque de Bayes, il est qualifié de bayésien.

4.1.5 Sélection de modèles

Les modèles et leurs paramètres ne peuvent pas être sélectionnés ensemble par le calcul direct de leurs vraisemblances. En effet, plus un modèle est complexe, plus il s'ajustera bien aux données. Partant de cette constatation, il est clair que pour un jeu fini de réalisations x de la variable donnée X, une sélection d'un modèle $\mathcal{M}$ et de ses paramètres θ, obtenue par la maximisation de la vraisemblance $p(x \mid \theta, \mathcal{M})$, aboutira inévitablement à la surestimation de la complexité du modèle.

Conscient de cet effet indésirable, le principe du rasoir d'Occam stipule que des modèles simples représentent mieux les données que des modèles complexes *(Jeffreys, 1939)*. Pour formaliser ce principe, un facteur multiplicatif est introduit dans le rapport de vraisemblance. Il est nommé facteur d'Occam.
L'existence de ce facteur est confortée par le fait qu'il se trouve être naturellement inclus dans le calcul de l'évidence bayésienne des modèles *(MacKay, 2003)*. En effet, la sélection de modèle est basée sur la maximisation de l'évidence obtenue par l'intégrale de marginalisation de l'Eq. 4.8. Or, pour beaucoup de problèmes, l'a posteriori $p(\theta \mid x, \mathcal{M}_i) \propto p(x \mid \theta, \mathcal{M}_i)p(\theta \mid \mathcal{M}_i)$ (Eq. 4.6) possède un pic important au niveau de l'estimateur $\widehat{\theta}_{MAP}$. Donc, l'évidence peut être évaluée en approximant l'a posteriori $p(\theta \mid x, \mathcal{M}_i)$ par une gaussienne autour du MAP. Cette approximation est obtenue par un développement du second ordre en série de Taylor du log de l'a posteriori autour de $\widehat{\theta}_{MAP}$. Elle conduit à l'approximation suivante

$$\begin{aligned} p(x \mid \mathcal{M}) &\simeq p(x \mid \widehat{\theta}_{MAP}, \mathcal{M}) \times p(\widehat{\theta}_{MAP} \mid \mathcal{M}) \, [det(H/2\pi)]^{-\frac{1}{2}} \\ Evidence &\simeq Vraisemblance \times Facteur\, d'Occam \end{aligned} \tag{4.15}$$

où $H = -\nabla^2 log\, p(\theta \mid x, \mathcal{M})$ est le laplacien du logarithme de l'a posteriori évalué en $\widehat{\theta}_{MAP}$.
Par l'emploi du logarithme et d'autres approximations, en notant par $\widehat{\theta}$ un estimateur obtenu par une procédure bayésienne (ML, MAP, ...), on peut montrer que le facteur d'Occam s' écrit pour n réalisations d'un modèle paramétrique régulier possédant k degré de liberté comme

$$log\, p(\widehat{\theta} \mid \mathcal{M}) - \frac{k}{2} log(\frac{n}{2\pi}) - log\, det^{-\frac{1}{2}}(\mathcal{J}) \tag{4.16}$$

où $det(\mathcal{J})$ est le déterminant de $\mathcal{J}$ dont les éléments

$$\mathcal{J}_{i,j} = E[\frac{\partial^2 log\, p(x|\Theta)}{\partial\theta_i \partial\theta_j}]. \tag{4.17}$$

constituent la matrice d'information de Fisher évaluée à $\widehat{\theta}$.
Pour sélectionner l'ordre d'un modèle paramétrique, *Schwarz (1978)* a recours à l'approximation de l'Eq. 4.16 dans laquelle il ne considère que le terme dominant $\frac{k}{2} log\, n$. Plus précisément, il utilise l'approximation d'ordre un du facteur d'Occam dans l'évidence pour dériver le critère d'information de Bayes (BIC)

$$BIC = -log\, p(x|\widehat{\Theta}) + \frac{k}{2} log\, n. \tag{4.18}$$

où k et n représentent respectivement l'ordre du modèle et le nombre de réalisations. La minimisation du critère BIC permet la sélection de l'ordre d'un modèle.

Jusqu'ici nous n'avons considéré que l'évidence pour la sélection de modèles. Or, le deuxième niveau d'inférence bayésienne (Eq. 4.7) fait apparaître un a priori sur les modèles $\mathcal{M}_i$. Donc, dans le cas où de l'information a priori sur les modèles est disponible, la maximisation de l'a posteriori

$$p(\mathcal{M}_i \mid x) \propto p(x \mid \mathcal{M}_i) p(\mathcal{M}_i)$$

permettra une meilleure sélection de modèles.

4.2 Mesure de l'information

Cette section introduit la notion d'information. Les mesures de l'information les plus connues sont l'entropie de Shannon, la divergence de Kullback-Leibler ou encore la complexité. La définition de ces notions d'information sera la base de concepts fondamentaux, tels que les distances probabilistes, l'indépendance statistique ou encore le codage. Ceux-ci seront utiles à la modélisation stochastique par des méthodes entropiques des signaux multidimensionnels qui sera abordée dans la section 4.3).

4.2.1 Entropie de Shannon

Plus l'incertitude concernant les réalisations d'une variable aléatoire est grande, plus l'information contenue dans celle-ci est importante. Sur cette base et en représentant par des probabilités le degré de plausibilité des réalisations de la variable aléatoire X, *Shannon (1949)* appela l'entropie de X, la moyenne d'information contenue dans la variable aléatoire $log(X)$. L'entropie continue (ou différentielle) de Shannon est définie par

$$H(X) = -\int p(x) \log p(x) dx. \tag{4.19}$$

On note que l'entropie est maximale lorsque la distribution de la variable aléatoire est uniforme, tandis que son entropie est nulle lorsque sa distribution est un Dirac (autrement dit si la variable est déterministe).

La définition de l'équation 4.19 s'applique dans le cas d'une variable aléatoire continue. Pour mesurer l'information d'une variable aléatoire discrète, Shannon dériva l'entropie discrète

$$H(X) = -\sum_{x_i} p(x_i)\log p(x_i). \tag{4.20}$$

Il est important de noter que l'entropie continue est différente et incomparable à l'entropie discrète malgré son aspect similaire. En particulier, il doit être remarqué que l'entropie continue n'est pas la valeur limite de l'entropie discrète obtenue en discrétisant l'axe réel x en intervalles infiniment petits[1]. Comme le précise *Jumarie (1990)*, l'entropie d'une variable continue est plus complète que celle d'une variable discrète dans le sens que pour cette dernière, l'entropie relative à la variabilité du treillis supportant la distribution n'est pas prise en compte. L'entropie totale de Shannon a été proposée par *Jumarie* pour unifier ces deux définitions.

Tout au long de cette thèse, l'entropie continue sera utilisée distinctement de l'entropie discrète. Plus particulièrement, l'entropie discrète de Shannon (Eq. 4.20) sera employée dans le contexte du théorème du codage de source de Shannon, qui souvent, constitue la base de la longueur de description en 2-parties du principe MDL. L'entropie continue (Eq. 4.19) sera utilisée pour l'ensemble des autres mesures et méthodes entropiques.

Dans une approche similaire, en accord avec l'approche bayésienne des probabilités, l'entropie conditionnelle de la variable X conditionnée par la variable Y est définie par

$$H(Y|X) = -\int\int p(x,y)\log p(y|x)dxdy. \tag{4.21}$$

L'entropie conjointe $H(Y,X)$ du couple de variable X et Y est définie par

$$H(Y,X) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{4.22}$$

4.2.2 Divergence de Kullback-Leibler et information mutuelle

Divergence de Kullback-Leibler La divergence de Kullback-Leibler est la généralisation de la mesure d'information de Shannon. La divergence est une fonction de deux distributions $p(x)$ et $q(x)$ caractérisant potentiellement une variable aléatoire X. La divergence

[1]En effet, un exemple très simple peut servir à illustrer la grande différence de nature entre l'entropie discrète et l'entropie continue : la variable continue X de distribution uniforme sur un intervalle $[a,b]$ et nulle en dehors de celui-ci, possède une entropie égale à $ln(b-a)$; si l'on discrétise l'intervalle $[a,b]$ en m sous-intervalles, on obtient une distribution discrète pour X, et l'entropie de cette variable discrétisée est égale à $ln(m)$; si maintenant on fait tendre m vers l'infini

$$\lim_{m\to\infty} ln(m) \neq ln(b-a),$$

et on ne retrouve pas la valeur de l'entropie continue.

entre les distributions $p(x)$ et $q(x)$ est définie par

$$D(p,q) = \int p(x)\log\frac{p(x)}{q(x)}dx. \tag{4.23}$$

Deux propriétés importantes de la divergence de Kullback-Leibler sont que $D(p,q) \geq 0$ et que $D(p,q) = 0$ si et seulement si $p(x) = p(y)$.
On peut interpréter la divergence de Kullback-Leibler comme la mesure de la quantité d'information donnée sur la vérité d'un modèle $p(x)$ lorsqu'il est comparé à un modèle alternatif $q(x)$ *(Kullback, 1966)*. On peut encore comprendre la divergence comme une entropie d'une distribution a posteriori $p(x)$ étant donnée les observations comparées à la distribution a priori $q(x)$. C'est pourquoi la divergence est aussi appelée entropie bayésienne. Un cas particulier est lorsque $q(x)$ est uniforme. Alors, la divergence est similaire à l'entropie de Shannon de X. Donc l'entropie de Shannon peut être considérée comme la quantité d'information dans un modèle $p(x)$ de X comparée au modèle d'entropie maximum - la distribution uniforme.
Cette mesure est à la base d'inférences entropiques basées sur le principe de l'information de discrimination minimum *(Topsoe, 1993)*, encore appelé discrimination minimum de Kullback (c.f. § 4.3.1).
Pour accéder à la notion de distance entre deux distributions, il faut employer la divergence symétrisée de Kullback-Leibler définie par

$$D_{sym}(p,q) = D(p,q) + D(q,p) = \int (p(x) - q(x))\log\frac{p(x)}{q(x)}dx. \tag{4.24}$$

Information mutuelle L'information mutuelle de Shannon $I(X,Y)$ est un cas particulier de la divergence de Kullback-Leibler. Elle représente la quantité d'information contenue dans la variable X sur une variable Y ou, parce que la mesure est symétrique, la quantité d'information contenue dans Y sur X. En d'autres termes, cette quantité représente la diminution de l'incertitude d'une variable par la connaissance d'une autre. Elle est définie comme la divergence de Kullback-Leibler entre la distribution conjointe $p(x,y)$ et le produit de ses distributions marginales $p(x)p(y)$

$$I(X,Y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy. \tag{4.25}$$

Employant les entropies conditionnelles ou encore l'entropie conjointe, elle peut se réécrire comme

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{4.26}$$

Deux propriétés importantes de l'information mutuelle sont que $I(X,Y) \geq 0$ et $I(X,Y) = 0$ si et seulement si X et Y sont statistiquement indépendantes, c'est à dire : $p(x,y) = p(x)p(y)$.

4.2.3 Longueur de code et complexité algorithmique

Les mesures entropiques précédentes sont fortement reliées à des mesures d'information basées sur le codage. Il existe en effet une relation étroite entre probabilités et longueur de code. Cette relation découle de l'inégalité de Kraft.

Inégalité de Kraft Introduisons d'abord quelques notations et définitions. Un codage $\mathcal{C}$ pour une collection de mots $x^n = \{x_1, ..., x_n\}$; $x_i \in \mathcal{A}$, où $\mathcal{A}$ représente un dictionnaire, résulte simplement d'une transformation de x^n vers une collection de mots codes. Un codage est appelé codage préfixe si aucun mot code est le préfixe d'un autre mot code, garantissant ainsi un décodage unique.
En général, il y a une correspondance entre la longueur de code issue d'un codage préfixe et les quantités $-log\, p(x_i)$ pour la distribution de probabilités p sur $\mathcal{A}$. En effet, une fonction retournant une valeur entière $L(.)$ et prenant pour argument un mot x_i, correspond à une longueur de code issue d'un codage préfixe binaire si et seulement si elle satisfait l'inégalité de Kraft :

$$\sum_{x_i} 2^{-L(x_i)} \leq 1, \tag{4.27}$$

voir *Cover et Thomas (1991)* pour la preuve. En conséquence, étant donné un codage préfixe $\mathcal{C}$ sur $\mathcal{A}$ avec la fonction de longueur L, on peut définir une distribution sur $\mathcal{A}$ comme ci dessous

$$p(x_i) = \frac{2^{-L(x_i)}}{\sum_{x_i} 2^{-L(x_i)}},\ \forall x_i \in \mathcal{A} \tag{4.28}$$

Inversement, pour toute distribution p sur $\mathcal{A}$, on peut trouver un codage préfixe avec une fonction de longueur $L(x_i) = \lceil -log\, p(x_i) \rceil$ où $\lceil . \rceil$ désigne l'arrondi à l'entier le plus proche.

Théorème du codage de source de Shannon Supposons maintenant que les éléments ou symboles de $\mathcal{A}$ soient générés avec une distribution connue p. Le théorème du codage de source de Shannon stipule que l'espérance de la longueur de code issue d'un codage $\mathcal{C}$ définie par

$$L_{\mathcal{C}} = -\sum_{x_i} p(x_i) L(x_i) \tag{4.29}$$

possède pour borne inférieure $H(X)$, l'entropie discrète de Shannon[2].

$$L_{\mathcal{C}} \geq H(X) = -\sum_{x_i} p(x_i) log\, p(x_i). \tag{4.30}$$

[2]Si l'on considère un codage binaire, la borne inférieure est l'entropie discrète définie avec un logarithme en base 2. Dans ce cas, l'unité de la longueur de code sera le bit.

L'égalité de l'Eq.4.30 est vérifiée si et seulement si $L(x_i) = -log\, p(x_i),\ \forall x_i$. Un tel codage engendre une longueur de code appelée longueur de code de Shannon-Fano.

Notons que l'on peut étendre le théorème de codage de Shannon d'une source à des variables aléatoires continues en utilisant l'entropie continue. On trouve ainsi que la borne inférieure de l'espérance de la longueur de code est constituée par l'entropie continue. Mais encore une fois, de par la nature différente de l'entropie continue et discrète (c.f.§ 4.2.1), il est important de noter que la borne continue est différente et incomparable à la borne discrète. Même si l'on se place dans le cas limite où l'on discrétise la distribution continue en intervalles infiniment petits, la borne continue et la borne discrète demeurent différentes.

La longueur de code issue du codage de Shannon quantifie l'information issue d'une source. Toutefois, cette notion de codage optimal est restreinte à la situation où l'on considère l'information moyenne communiquée par les réalisations de la source.

Complexité algorithmique de Kolmogorov En contraste avec la théorie de l'information qui ne considère pas l'information absolue contenue dans chaque réalisation d'une source prise individuellement, *Kolmogorov (1965)* introduisit la complexité intrinsèque d'une série. La complexité algorithmique de Kolmogorov $K(X)$ est égale à une constante près à la longueur minimale du programme p à décoder par une machine Ψ pour reproduire la série x^n. La complexité de Kolmogorov recherche une longueur moyenne de codage d'un mot proche de l'entropie, mais de plus, les régularités de la série particulière sont toutes prises en compte. Le programme p est donc aléatoire au sens de *Martin-Löf (1966)*. Malheureusement, cette quantité n'est pas calculable et pour des applications pratiques des approximations sont utilisées.
La longueur de code de Shannon-Fano et la complexité de Kolmogorov sont dans une certaine mesure équivalentes. Les équivalences reposent sur la validité de la substitution des probabilités de l'Eq. 4.30 par des modèles universels. Ces derniers seront définis dans la section 4.3.4 qui abordera les relations entre ces deux types de mesure de l'information.

4.3 Méthodes entropiques pour l'extraction d'information

Les méthodes entropiques se basent sur les mesures d'information précédemment présentées. Elles permettent l'inférence de modèles et de paramètres et servent à la décision.

4.3.1 Discrimination minimum de Kullback

Le principe de la discrimination minimum de Kullback est un principe qui s'applique pour la sélection d'un modèle $p(x)$ lorsqu'un a priori $q(x)$ est disponible et que l'on a connaissance d'un jeu de contraintes que doit respecter $p(x)$. Le principe affirme que parmi

les distributions de probabilités satisfaisant les contraintes, on doit choisir la distribution $\widehat{p}(x)$ qui minimise la divergence i.e.

$$D(\widehat{p}, q) = \min_{p} D(p, q) \tag{4.31}$$

où le minimum est pris parmi le jeu de $p(x)$ satisfaisant les contraintes. Ces dernières peuvent être formulées par l'utilisation de facteurs de Lagrange.
Par exemple, si l'on connaît seulement la distribution a priori $q(x)$ d'une variable aléatoire X, on se doit de choisir une distribution $\widehat{p}(x)$ égale à l'a priori $q(x)$ car cette distribution minimise la divergence $D(p, q)$. Si toutefois une autre information est disponible, comme par exemple la connaissance de la moyenne de la distribution, i.e., si $E[X] = m_x$, on ne peut plus choisir $\widehat{p}(x) = q(x)$ car la contraintes de moyenne ne sera peut être pas satisfaite. Il faudra choisir la distribution $\widehat{p}(x)$ qui minimise $D(p, q)$ et qui ait pour moyenne m_x. Si la distribution a priori $q(x)$ n'est pas connue, alors on prendra un $q(x)$ uniforme dans le cas discret et gaussien dans le cas continu et on choisira un $\widehat{p}(x)$, satisfaisant les contraintes et qui minimise la divergence. Ce cas particulier de la discrimination minimum de Kullback est le principe du maximum d'entropie de *Jaynes (1957)*.
La minimisation de la divergence de Kullback-Leibler sera employée pour l'inférence de trajectoires dans la section 6.2.1.

4.3.2 Critère d'information de Akaike

D'après le paradigme du rasoir d'Occam, la sélection de l'ordre d'un modèle ne peut pas se baser sur l'estimateur du maximum du log de la vraisemblance (c.f. § 4.1.4). *Akaike (1973)*, sur la base de la divergence de Kullback-Leiber, a développé le critère d'information de Akaike (AIC) pour la sélection non biaisée de l'ordre d'un modèle.
On considère que $\mathcal{M}_0$ associé à un certain vecteur de paramètres θ_0 est le vrai modèle génératif de réalisations $x^n = \{x_1, ..., x_n\}$ de la variable aléatoire ou du processus stochastique X. On note la vraisemblance du vecteur de paramètres estimé par ML $p(x^n \mid \widehat{\theta}_{ML})$ et la vraisemblance du véritable vecteur de paramètres $p(x^n \mid \theta_0)$. La divergence entre les distributions $p(x^n|\theta_0)$ et $p(x^n|\widehat{\theta}_{ML})$ s'écrit

$$D\big(p(x^n|\theta_0), p(x^n|\widehat{\theta}_{ML})\big) = \int p(x^n|\theta_0) log \frac{p(x^n|\theta_0)}{p(x^n|\widehat{\theta}_{ML})} dx^n \tag{4.32}$$

Cette mesure représente l'information perdue dans la fonction du log de vraisemblance quand $\widehat{\theta}_{ML}$ est utilisé pour approximer θ_0. L'Eq. 4.32 peut se réécrire

$$\begin{aligned} D\big(p(x^n|\theta_0), p(x^n|\widehat{\theta}_{ML})\big) &= E_x[log\, p(x^n|\theta_0)] - E_x[log\, p(x^n|\widehat{\theta}_{ML})] & (4.33) \\ &= C - E_x[log\, p(x^n|\widehat{\theta}_{ML})] & (4.34) \end{aligned}$$

où C est une constante et $E_x[.]$ est l'espérance calculée avec les réalisations x^n.
$C - D\big(p(x^n|\theta_0), p(x^n|\widehat{\theta}_{ML})\big) = E_x[log\, p(x^n|\widehat{\theta}_{ML})]$ représente une divergence relative entre les deux distributions. Pour rendre la mesure indépendante des réalisations particulières

x^n, on applique des deux côtés de l'expression l'espérance par rapport à des réalisations indépendantes x^{n*}

$$C - E_{x*}\big[D\big(p(x^n|\theta_0), p(x^n|\widehat{\theta}_{ML})\big)\big] = E_{x*}\big[E_x[log\, p(x^n|\widehat{\theta}_{ML})]\big] \tag{4.35}$$

La découverte d'Akaike d'une relation entre le maximum du log de vraisemblance et l'espérance de la divergence relative (ou la moyenne de l'espérance du maximum du log de vraisemblance) permit des avancées pratiques et théoriques majeures dans la sélection de modèles. La relation établit que, sous la condition que la distribution soit différentiable en Θ, le biais entre ces deux quantités est égal à l'ordre du modèle k *(Sakamoto et al., 1986)*

$$E_{x*}\big[E_x[log\, p(x^n|\widehat{\theta}_{ML})]\big] = -log\, p(x^n|\widehat{\theta}_{ML}) + k. \tag{4.36}$$

Pour des raisons historiques Akaike définit son critère avec un facteur 2. Ce critère s'écrit

$$AIC = -2log\, p(x^n|\widehat{\theta}_{ML}) + 2k. \tag{4.37}$$

L'ordre d'un modèle peut être évalué par minimisation du critère AIC pour un nombre n assez grand de réalisations. Cependant, AIC est asymptotiquement optimal et sélectionne le modèle s'accordant le plus aux données et non le modèle véritable pour un jeu fini de réalisations.

4.3.3 Principe de la longueur minimale de description (MDL)

Dans la section 4.2.3, nous avons mis en évidence le rapport entre les longueurs de code et les distributions de probabilités. Dans le contexte de la compression de données, la motivation est de trouver un bon codage des données, en employant des modèles pour construire ces codes s'approchant le plus de la véritable distribution de probabilités. Dans un contexte différent, le principe MDL utilise la minimisation de la longueur de description des données comme critère pour la sélection d'une classe de modèles ou d'un modèle particulier. Cette idée a d'abord été formulée par *Solomonoff (1964)* pour la prédiction de séquences utilisant la description efficace la plus courte.

Cette approche a inspiré plus tard Rissanen *(Rissanen 1978, Rissanen 1986)* pour dériver le principe MDL :

Pour un certain jeu de réalisations et une collection finie de modèles, le principe MDL sélectionne le modèle engendrant la longueur minimale de description des données.

La validité du principe MDL dépend des propriétés de la longueur de description employée ou plus précisément des propriétés du schéma de codage sous-jacent. Pour formaliser ces propriétés, *Rissanen (1986)* dériva la borne inférieure de la redondance (point wise lower bound) puis *Clarke et Barron (1990)* dérivèrent la borne inférieure de la redondance minimax du code avec lequel les données sont encodées pour une classe de modèles donnée. Des schémas de codage aboutissant à des longueurs de description dont la redondance vérifient ces bornes inférieures seront appelés codages universels. Ils produiront des critères

MDL valides pour la sélection de modèles *(Hansen et Yu, 2001)*.
Dans un premier temps, les deux types de bornes inférieures seront présentées puis dans un deuxième temps, on présentera la longueur de description en 2-parties et la complexité d'information stochastique (SIC) qui sont des schémas de codage universels atteignant ces bornes.

4.3.3.1 Bornes de la redondance d'un codage

Dans ce paragraphe, on se placera dans le contexte de classes de modèles paramétriques. Toutefois, ces bornes inférieures de la notion de redondance du codage peuvent être généralisées pour certaines classes de modèles non-paramétriques.
Définissons d'abord ce qu'est l'espérance de la redondance d'un codage. Considérons une classe de modèles paramétriques

$$\mathcal{M} = \{p(x^n \mid \theta), \theta \in \mathbf{R}^k\}$$

constituée d'une collection de modèles paramétriques d'ordre variable k. Notons par $\mathcal{M}_0$ le modèle génératif des réalisations $x^n = \{x_1, ..., x_n\}$ de la variable aléatoire ou du processus stochastique X. Notons par $E_0\{.\}$ l'espérance relative à la distribution de probabilités $p_0(x^n) = p(x^n \mid \Theta_0)$ donnée par le vrai modèle $\mathcal{M}_0$. En utilisant ces notations, l'entropie différentielle de X s'écrit

$$H(X) = -E_0\{log\, p_{_0}(x^n)\}.$$

Pour toute distribution de probabilités (ou codage préfixe) $q(x^n)$, la divergence de Kullback-Leibler entre les distributions p_0 et q est donnée par

$$\begin{aligned} D(p_{_0}(x^n), q(x^n)) &= E_0\{log\frac{p_{_0}(x^n)}{q(x^n)}\} \\ &= E_0\{-log\, q(x^n) - (-log\, p_{_0}(x^n))\} \\ &= R_n(p_{_0}, q) \end{aligned} \tag{4.38}$$

$R_n(p_{_0}, q)$ représente l'espérance de la longueur de code supplémentaire nécessaire à l'encodage des n réalisations x^n de la variable X^n utilisant q plutôt que le schéma de codage optimal $p_{_0}$. Dans la théorie du codage, R_n est appelé l'espérance de la redondance du codage q. Définir une longueur de description valide des données x^n basée sur les modèles de la classe $\mathcal{M}$, revient à trouver une distribution q qui atteigne la plus petite redondance possible parmi tous les membres de $\mathcal{M}$.

– Pour concrétiser cette idée, une première borne inférieure de la redondance est dérivée par *Rissanen (1986)*. En supposant qu'un estimateur $\widehat{\theta}$ se calcule avec une précision relative à la taille des données $\delta_n = 1/\sqrt{n}$, pour toute distribution q, Rissanen trouve que

$$\lim_{n\to\infty} \frac{R_n(p_{_0}, q)}{(k/2)log\, n} \geq 1. \tag{4.39}$$

où k représente la dimension du vecteur de paramètres θ. Cette inégalité stipule que, ignorant la véritable distribution p_0, quelque soit le codage préfixe utilisé, on aura besoin d'une longueur de code supplémentaire au minimum égale à $(k/2)log\, n$ pour encoder x^n.

– Une borne inférieure de la redondance minimax est ensuite définie. Elle s'exprime par

$$R_n^+ = \min_q \sup_\theta R_n(p_0, q) \quad = \quad \frac{k}{2} log \frac{n}{2\pi} + log\sqrt{det(\mathcal{J})} - log\, p(\theta) \qquad (4.40)$$

où $\mathcal{J}$ est la matrice d'information de Ficher définie par l'Eq. 4.17. Cette borne n'est autre que l'approximation du facteur d'Occam (Eq. 4.16). La littérature sur MDL rejette les probabilités a priori subjectives, telles qu'elles peuvent apparaître dans la théorie de Bayes. Rissanen choisit donc l'a priori non-informatif de Jeffrey

$$p(\theta) = \frac{\sqrt{det(\mathcal{J})}}{\int \sqrt{det(\mathcal{J})} d\theta} \qquad (4.41)$$

pour finalement trouver la borne inférieure de la redondance minimax

$$R_n^+ \quad = \quad \frac{k}{2} log \frac{n}{2\pi} + log \int \sqrt{det(\mathcal{J})} d\theta + o(1) \qquad (4.42)$$

On trouve donc que la première borne inférieure $\min_q R_n(p_0, q)$ est l'approximation à l'ordre 1 de la borne inférieure de la redondance minimax. Le principe MDL affirme que tout codage atteignant cette approximation à l'ordre 1 de l'espérance de la redondance, constitue un codage valide pour représenter la longueur de description pour la sélection de modèles. Un tel codage sera appelé "codage universel" noté $CodeUniv(.)$.

On note que le théorème du codage de Shannon (c.f. § 4.2.3) quantifie la meilleure espérance de longueur de code quand le véritable modèle de la source est connu. La borne inférieure de Rissanen étend ce résultat au cas où l'on sait seulement que le véritable modèle appartient à une classe de modèles. Le critère AIC (c.f. § 4.3.2) se base aussi sur l'analyse de la divergence de Kullback-Leibler ou la redondance $R_n(p_0, q)$. Mais pour AIC la distribution q est remplacée par la distribution $p_{\widehat{\theta}_{ML}}$ associée à l'estimateur du maximum de vraisemblance $\widehat{\theta}_{ML}$. Sous des hypothèses standards abordées dans la partie 4.3.2, Akaike trouva que $R_n(p_0, p_{\widehat{\theta}_{ML}})$ a une moyenne limite égale à $k/2$ due au processus d'estimation paramétrique. Cette moyenne limite constitue la moitié de la correction AIC, l'autre moitié étant due à l'introduction d'une espérance de $R_n(p_0, p_{\widehat{\theta}_{ML}})$ *(Sakamoto et al., 1986)*.

4.3.3.2 Codages universels pour la sélection de modèles

Dans cette section, basées sur des classes de modèles, on introduit des codages universels de données qui constituent des critères MDL valides : les codages ont une redondance

atteignant la borne inférieure définie précédemment. On présentera la complexité d'information stochastique et des longueurs de description en 2-parties qui seront, comme on le verra, respectivement très proches de l'évidence bayésienne et du critère BIC.

Complexité d'information stochastique En se basant sur la borne inférieure de la redondance minimax R_n^+, Rissanen dériva un critère pour les classes de modèles paramétriques régulières qu'il nomma complexité d'information stochastique. Comme *MacKay (1992)* le souligne, cette quantité est équivalente à l'évidence bayésienne (Eq. 4.15)

$$SIC(x^n) = -log\, p(x^n \mid \mathcal{M}_i) = -log \int p(x^n \mid \theta, \mathcal{M}_i) p(\theta \mid \mathcal{M}_i) d\theta \tag{4.43}$$

et permet donc la sélection d'une classe de modèle $\mathcal{M}_i$. Cette quantité est comme l'évidence bayésienne difficilement calculable. Une méthode populaire pour son évaluation est l'utilisation des méthodes de Monte Carlo impliquant les chaînes de Markov (*George et McCulloch, 1997)*. Face à l'énorme coût calculatoire, on emploie souvent des approximations pour l'estimation de l'ordre d'un modèle paramétrique. En notant par $\widehat{\theta}$ un estimateur des paramètres θ du modèle $\mathcal{M}_i$ obtenu par une procédure bayésienne (ML, MAP, ...), la complexité d'information stochastique est approximée par

$$SIC(x^n) \approx -log\, p(x^n \mid \widehat{\theta}, \mathcal{M}_i) + \frac{k}{2} log \frac{n}{2\pi} + log \int \sqrt{det(\mathcal{J})} d\theta \tag{4.44}$$

C'est l'approximation à l'ordre 2 (Eq. 4.16) de l'évidence bayésienne où la distribution $p(\theta \mid \mathcal{M}_i)$ est un a priori de Jeffrey.

Longueur de description en 2-parties De même, en se basant sur la borne inférieure $\min_q R_n(p_0, q)$, on dérive la longueur de description en 2-parties. Cette longueur de description est un schéma de codage universel qui présente des facilités d'implémentation. En effet, cette longueur de description se décompose comme son nom l'indique en 2 parties :

- **Partie 1 :** Longueur de description $-CodeUniv(\mathcal{M}_i)$ nécessaire pour décrire le modèle $\mathcal{M}_i$ et notamment ses paramètres estimés $\widehat{\theta}$. Cette quantité est représentative de la complexité du modèle. Dans le formalisme MDL, on emploie une distribution a priori non-informative sur les modèles, telle que la distribution uniforme, pour dériver ce codage universel.

- **Partie 2 :** Longueur de description $-CodeUniv(x^n \mid \mathcal{M}_i)$ nécessaire pour coder les données x^n connaissant le modèle $\mathcal{M}_i$ et ses paramètres estimées $\widehat{\theta}$.

La longueur de description en 2-parties s'exprime donc par

$$L_{2P}(x^n) = -CodeUniv(\mathcal{M}_i) - CodeUniv(x^n \mid \mathcal{M}_i) \tag{4.45}$$

Analysons la première partie de la longueur de description dans le cas de familles de modèles paramétriques. Notons, que pour sélectionner l'ordre k d'un modèle paramétrique,

le codage universel du modèle se résume au codage du vecteur $\widehat{\theta}$ constitué par les k paramètres estimés. On emploiera souvent une distribution continue uniforme sur l'espace de paramètres, pour coder les k paramètres du modèle. Pour se ramener dans le cas discret et appliquer le théorème du codage de source de Shannon, on discrétise la distribution continue uniforme de chaque paramètre en intervalles régulier de taille $\delta_1 = 1/(\sqrt{n})$. En notant par $p([\widehat{\theta}]_\delta)$ la distribution uniforme discrétisée, on obtient ainsi une première longueur de description égale à

$$-\,CodeUniv(\mathcal{M}_i) = -log\, p([\widehat{\theta}]_{\delta_1}) - k\, log\, \delta_1 \tag{4.46}$$

qui est souvent tronquée en ne considérant que le terme du premier ordre

$$\begin{aligned} -\,CodeUniv(\mathcal{M}_i) &= -log\, p([\widehat{\theta}]_{\delta_1}) + \frac{k}{2} log(n) \\ &= \frac{k}{2} log(n) + o(1). \end{aligned} \tag{4.47}$$

Analysons la deuxième partie de la longueur de description dans le cas de familles de modèles paramétriques régulières. Comme l'on doit se placer dans le cas discret pour l'application du théorème du codage de source de Shannon, on discrétise la distribution continue de vraisemblance avec une précision égale à δ_2 pour former ainsi une longueur de description égale à

$$-\,CodeUniv(x^n \mid \mathcal{M}_i) = -log\, p([x^n]_{\delta_2} \mid [\widehat{\theta}]_{\delta_1}, \mathcal{M}_i) - n\, log\delta_2. \tag{4.48}$$

Or la différence entre l'opposé du logarithme de la vraisemblance continue et discrétisée est négligeable pour des distributions suffisamment lisses, comme c'est le cas pour des familles de modèles paramétriques régulières. De plus le terme $n\, log\delta_2$ est constant indépendamment du modèle $\mathcal{M}_i$. C'est pourquoi dans la littérature, souvent la longueur de description $-CodeUniv(x^n \mid \mathcal{M}_i)$ est simplement approximée par l'opposé du logarithme de vraisemblance $-log\, p(x^n \mid \widehat{\theta}, \mathcal{M}_i)$ des données calculée pour l'estimateur $\widehat{\theta}$ et le modèle $\mathcal{M}_i$.

En employant l'approximation de la deuxième partie dans l'Eq. 4.45, la longueur de description en 2-parties devient

$$L_{2P}(x^n) = -log\, p(x^n \mid \widehat{\theta}, \mathcal{M}_i) - CodeUniv(\mathcal{M}_i). \tag{4.49}$$

L'approximation de l'équation 4.47 rapporté dans l'Eq. 4.49, démontre l'équivalence entre $L_{2P}(x^n)$ et le critère BIC. On remarque que $L_{2P}(x^n)$ n'est autre que l'approximation à l'ordre 1 de la $SIC(x^n)$. En effet, on peut prouver que

$$\begin{aligned} \frac{k}{2} log \frac{n}{2\pi} + log \int \sqrt{det(\mathcal{J})} d\theta &= \frac{1}{2} log\, det(n\, \mathcal{J})(1 + o(1)) \\ &= \frac{k}{2} log\, n(1 + o(1)). \end{aligned} \tag{4.50}$$

Dans un contexte plus général, *Clarke et Barron (1990)* ont prouvé qu'une longueur de description en 2-parties est une approximation du premier ordre de SIC.

Dans l'idéal, le principe MDL sélectionne une longueur de description en deux parties minimum issue de la complexité algorithmique de Kolmogorov

$$L_{2P}^{Ideal}(x^n) \quad = \quad K(x^n \mid \mathcal{M}_i) + K(\mathcal{M}_i). \tag{4.51}$$

Mais cette quantité est difficilement calculable, et en pratique on substitue ces quantités par les codages universels présentés ci-dessus.

Longueur minimale de message Bien que des résultats similaires soient obtenus en pratique, le principe MDL diffère par quelques points *Wallace et Dowe (1999)* du principe de la longueur minimale de message (MML) développé indépendamment du principe MDL par *Wallace et Boulton (1968)*.
Premièrement, la philosophie du principe MML est de s'appuyer sur le maximum de connaissances a priori disponibles, contrairement aux développements MDL qui considèrent des a priori non-informatifs. Dans ce sens, le formalisme MML permet une interprétation encore plus similaire au deuxième niveau de l'inférence bayésienne.
Deuxièmement, le formalisme MML se concentre principalement sur l'emploi d'un codage en 2-parties, similaire à la longueur de description en 2-parties MDL mais avec des a priori sur les modèles $\mathcal{M}_i$ introduits explicitement.
Une autre différence entre l'approche MDL et MML est que la première sélectionne une classe de modèles tandis que la dernière sélectionne un modèle particulier. Mais cette affirmation est controversée. En effet, en théorie, la complexité d'information stochastique de MDL, sélectionne une classe de modèle de façon analogue à l'évidence. En pratique, on a recours à des approximations de l'évidence, telles que $L_{2P}(x^n)$, qui sélectionne un modèle complètement spécifié (paramètres estimés) et non une classe de modèles. En conséquence, les approches sont en pratique très proches.

Nous avons vu que les méthodes entropiques pour la sélection de modèles se basent toutes sur la divergence de Kullback-Leibler impliquant le modèle véritable générateur des données ou bien un modèle a priori disponible. Cette divergence entre modèles est minimisée pour sélectionner un modèle par discrimination minimum de Kullback ou pour obtenir un codage universel utilisé dans le formalisme MDL pour la sélection de modèles. Notons que ces principes peuvent être vus sous un angle commun : en se basant sur la divergence (ou l'espérance de la divergence) de Kullback-Leibler entre un modèle à évaluer et un modèle a priori (véritable ou non), ces principes infèrent des probabilités associées à chaque modèle candidat.

4.3.4 Equivalences entre les méthodes entropiques et bayésiennes

L'analogie entre la minimisation de la divergence de Kullback-Leibler, pierre angulaire des méthodes entropiques parfois appelée entropie bayésienne, et du formalisme bayésien est très forte. En effet, comme le souligne la section 4.2.2, on peut interpréter la divergence comme l'extraction d'information a posteriori à partir de connaissances a priori. Cette idée est la base du principe de discrimination minimum de Kullback. Mais l'analogie avec l'inférence bayésienne la plus complète se trouve dans le principe MDL. Les similarités sont

telles que la communauté scientifique juge parfois les résultats de ces théories similaires comme l'illustre *MacKay, (1992)* dans sa thèse à la page 17 :

‘ *‘Although some of the earliest work on complex model comparison involved the MDL framework, MDL has no apparent advantages, and in my work I approximate the evidence directly”*

Toutefois, ces principes d'inférences présentent quelques différences. Dans le chapitre 4.3, nous avons déjà établi certaines équivalences entre le formalisme bayésien et le principe MDL. L'objet de cette section est de clarifier et de résumer ces équivalences dans la sélection de modèles.

Premièrement, rappelons l'équivalence totale entre l'évidence bayésienne et la complexité d'information stochastique qu'exprime MacKay dans sa thèse. Ce sont deux dénominations de la même quantité (au logarithme près) issues de deux approches différentes : cadre bayésien et théorie du codage. Cette quantité s'exprime par

$$-\log p(x^n \mid \mathcal{M}_i) = -log \int p(x^n \mid \theta, \mathcal{M}_i) p(\theta \mid \mathcal{M}_i) d\theta,$$

et sa minimisation permet la sélection d'une classe de modèles (c.f. § 4.1.5 et c.f. § 4.3.3). Mais, parce que cette quantité présente une complexité calculatoire importante, elle est souvent remplacée par des approximations du premier et du second ordre.
Pour le second ordre, cette quantité est approximée identiquement dans le cadre bayésien et pour le principe MDL, à l'exception près que le contexte MDL remplace l'a priori sur les modèles par un a priori non-informatif de Jeffrey.
En ce qui concerne les approximations du premier ordre, pour des classes de modèles paramétriques, on obtient dans le formalisme bayésien le critère BIC qui est, comme précisé dans la dernière partie, équivalent à une forme de la longueur de description en 2-parties du principe MDL. D'une façon plus générale, on peut prouver qu'une longueur de description en 2-parties est une approximation du premier ordre de l'évidence bayésienne (ou de SIC). Cette approximation, présentant des facilités d'implémentation, constitue une alternative intéressante au calcul coûteux de l'évidence.

Deuxièmement, explicitons l'équivalence entre le principe MDL (Eq. 4.49) employant une longueur de description en 2-parties et le deuxième niveau d'inférence bayésienne utilisant un certain type de distribution a priori.
La sélection bayésienne du modèle du maximum a posteriori choisit le modèle $\mathcal{M}_i$ qui minimise l'expression

$$-\log p(\mathcal{M}_i \mid x^n) \propto -log\, p(x^n \mid \mathcal{M}_i) - \log p(\mathcal{M}_i).$$

Dans le principe MDL/MML, ces 2 quantités sont remplacées par des longueurs de code issues d'un codage universel. La première longueur de description est l'opposé du logarithme de la vraisemblance du modèle $-log\, p(x^n \mid \widehat{\theta}, \mathcal{M}_i)$ tandis que la deuxième longueur de description est le codage universel $-CodeUniv(\mathcal{M}_i)$ du modèle et de ses paramètres. Donc, contrairement a l'approche bayésienne, on n'introduit pas de connaissance a priori $p(\mathcal{M}_i)$ sur le modèle. Pour qu'une équivalence existe, il faut que cette distribution a

priori du formalisme bayésien soit non-informative. Dans ce dernier cas, on retrouve les équivalences entre la longueur de description en 2-parties et l'évidence bayésienne décrites dans le paragraphe précédent.

Enfin, montrons dans quelle mesure le principe MDL idéal (Eq. 4.51) employant une longueur de description en 2-parties est équivalent au deuxième niveau d'inférence bayésienne. Dans le principe idéal MDL, ces 2 logarithmes négatifs de probabilités sont substitués par la complexité algorithmique de Kolmogorov :

$$L_{2P}^{Ideal}(x^n) \quad = \quad K(x^n \mid \mathcal{M}_i) + K(\mathcal{M}_i).$$

Le problème est donc de savoir quand cette substitution est valide. *Vitanyi et Ming (2000)* montrent que dans certains cas simples où la collection de modèles est finie, le principe MDL idéal et la sélection bayésienne de modèles utilisant des a priori universels sur les modèles coïncident. Cependant, dans un contexte plus général, le principe idéal MDL et le modèle MAP bayésien peuvent diverger dû à la distinction entre la longueur de code de Shannon et la complexité algorithmique de Kolmogorov. Les auteurs quantifient cette divergence par une inégalité fondamentale.

En se basant sur les fondements théoriques solides du principe MDL et des rapports étroits avec le formalisme bayésien, on développera dans la section 6.1.3 une longueur de description en 2-parties, possédant une complexité calculatoire réduite, pour approximer l'évidence bayésienne et ainsi, inférer la complexité d'un modèle de mélange de gaussiennes représentatif d'un espace de caractéristiques. Nous verrons dans la section 4.5 que les performances de la modélisation seront liées à l'efficacité de l'algorithme d'optimisation employé. Cette modélisation présentera des similarités avec le programme *Snob* de classification par MML de *Wallace et Dowe (1994)*.

4.4 Modèles stochastiques pour les signaux multidimensionnels

Notre approche étant bayésienne, elle se base sur des modèles stochastiques, introduits notamment au travers des distributions a priori. Ces modèles doivent donc être suffisamment puissants pour représenter efficacement des signaux multidimensionnels. Dans cette section, on se basera sur les champs aléatoires de Gibbs, qui sont des modèles stochastiques appropriés à l'analyse de signaux multidimensionnels, pour dériver des modèles de processus stochastiques spatiaux. On présentera aussi des modèles et des mesures de similarités de graphes, représentatifs de processus spatio-temporels. On dérivera ensuite différents modèles pour des variables aléatoires multidimensionnelles qui donneront lieu à des analyses de structures linéaires ou non-linéaires.

4.4.1 Modèles de processus stochastiques : les champs aléatoires de Gibbs

Voisinages et systèmes de cliques Pour toute variable $X_i \in X^n = \{X_0, ..., X_n\}$, un voisinage N_i peut être défini par

$$N_i = \{j\} \text{ tel que } \begin{cases} i \notin N_i \\ j \in N_i \iff i \in N_j \end{cases}$$

Un voisinage est associé à un certain système $\mathcal{C}$ de cliques $\mathcal{C}_k$

$$\mathcal{C} = \{\mathcal{C}_k\},$$

chacune constituée de k variables aléatoires qui caractérisent un vecteur de paramètres

$$\theta = \{\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, ...\}$$

qui exprime les interactions locales agissant dans les cliques $\mathcal{C}_k$.
Notons que les cliques sont des dépendances entre variables possédant des relations spatiales et/ou temporelles ou de toute autre nature.

Energies et Potentiels Les interactions locales, c'est à dire des interactions qui agissent entre des éléments d'une clique, peuvent maintenant servir à exprimer des énergies et des potentiels. La fonction d'énergie de la variable conjointe s'exprime comme une somme, sur toutes les cliques c du modèle, de potentiels V_c correspondant

$$U(x^n \mid \theta) = \sum_{c \in \mathcal{C}} V_c(x^n \mid \theta). \tag{4.52}$$

La fonction d'énergie d'une variable seule s'exprime comme la somme de toutes les énergies des cliques qui incluent la variable. Elle s'écrit

$$U_i(x_i \mid \theta) = \sum_{c \in \mathcal{C}, i \in c} V_c(x_i \mid \theta). \tag{4.53}$$

Les champs aléatoires de Gibbs Pour un champ aléatoire de Gibbs, la distribution de probabilités peut être exprimée en terme de potentiel par

$$p(x^n \mid \theta) = \frac{1}{\sum_{x_i} \exp(-U(x_i \mid \theta))} \exp(-\sum_{c \in \mathcal{C}} V_c(x^n \mid \theta)). \tag{4.54}$$

Le dénominateur est appelé la fonction de partition de l'espace de configuration. Cette dernière est souvent difficile à évaluer du fait du nombre important n d'éléments de $\boldsymbol{X}^n$.
Pour formaliser les champs aléatoires de Gibbs, on a au préalable défini un système de cliques supportant des fonctions potentielles liant les variables aléatoires entre elles. On engendre ainsi une collection de noeuds et d'arcs qui engendrent une représentation du processus sous forme de graphe *(Geman et Geman, 1984)*.

Les champs aléatoires de Markov Pour un tel champ aléatoire, la distribution de probabilités d'une variable aléatoire peut être exprimée en fonction d'un nombre de voisins limité : pour $\overline{x_i} = \{x_j\}_{j \neq i}$, on a la propriété de Markov

$$p(x_i \mid \overline{x_i}) = p(x_i \mid x_j \in N_i). \tag{4.55}$$

Cette propriété a été présentée auparavant (Eq. 4.3) dans le cadre des réseaux bayésiens.

Le théorème de Hammersley-Clifford Celui ci établit qu'il y a une équivalence totale entre les champs aléatoire de Markov et de Gibbs. Il en découle que l'on peut avoir une vision locale des champs aléatoires de Gibbs définie par

$$p(x_i \mid x_j \in N_i, \theta) = \frac{1}{\sum_{x_i, x_j} \exp(-U(x_i \mid x_j \in N_i, \theta))} \exp(-U(x_i \mid x_j \in N_i, \theta)). \tag{4.56}$$

Comme tout champ aléatoire peut être modélisé comme un champ aléatoire de Markov avec un voisinage suffisamment grand, tout champ aléatoire peut être modélisé comme un champ aléatoire de Gibbs avec un système de clique suffisamment grand.

Des champs aléatoires de Gibbs seront utilisés pour l'extraction de caractéristiques de texture dans la section 6.1.1.1. Ces modèles sont aussi adaptés à la modélisation de processus temporels. Mais, plusieurs contraintes, comme l'échantillonnage temporel irrégulier ou un nombre insuffisant d'échantillons pour l'estimation, empêchent leur application directe. Toutefois, l'information du processus temporel sera considérée à travers l'espace multi-temporel des caractéristiques (c.f. § 2.4) et, a un plus haut niveau sémantique (c.f. § 4.4.2), au travers d'un modèle de similarité entre graphes présenté dans la section suivante.

4.4.2 Modèles et mesures de similarité entre graphes

Nous avons vu que pour les champs de Gibbs, les graphes peuvent être des modèles pour les processus stochastiques spatio-temporels. Les variables aléatoires $\boldsymbol{X^n} = \{X_1, ..., X_n\}$ constituent une collection de noeuds $N(\mathcal{G})$ du graphe $\mathcal{G}$, et les cliques $\mathcal{C} = \{\mathcal{C}_k\}$ constituent une collection d'arcs $A(\mathcal{G})$. Par des fonctions potentielles construites sur des connaissances a priori, on peut définir les probabilités $\{p(x_i \mid x_j \in N_i, \theta)\}$ de chaque noeud x_i connaissant son voisinage N_i *(Tupin et al., 1998 ; Quartulli et Datcu 2003).*

Intéressons nous maintenant aux mesures similarités de graphes envisageables. A partir d'une modélisation stochastique des processus aléatoires par les champs de Gibbs, une mesure de similarité intéressante entre graphes est la distance euclidienne entre les paramètres estimés $\widehat{\theta}$. On a alors accès à une caractéristique des graphes dans un repère orthonormé.
Pour une similarité évaluée en "temps réel", une alternative présentant un coût calculatoire moins élevé que l'évaluation d'une distance probabiliste par les champs aléatoires de Gibbs, est l'appariement inexact de *Bunke et Allermann (1983)*, qui mesure une similarité déterministe entre graphes. L'idée de l'appariement inexact entre deux graphes est de transformer un graphe en l'autre en associant un coût élémentaire à chaque distorsion

engendrée. On associera donc à chaque transformation une fonction de coût global, somme de tous les coûts élémentaires la composant. On minimisera, parmi toutes les combinaisons de transformations élémentaires possibles, la fonction de coût globale pour aboutir à la mesure de similarité de l'appariement inexact.
Autrement dit, en notant respectivement par $\nu_1 = \{\nu_i^1\}$ et par $\nu_2 = \{\nu_i^2\}$ les collections de noeuds des graphes $\mathcal{G}_1$ et $\mathcal{G}_2$, et notant par $\lambda = \{\lambda_i\}$ une collection supplémentaire de noeuds, on définit une transformation globale $\mathcal{F} = \{f\}$, composée par une certaine combinaison de transformations élémentaires $f : \nu^1 \rightarrow \nu^{2\lambda} = \nu^2 \cup \lambda$. Une fonction de coût $C(f(\nu_i^1) = \nu_j^{2\lambda})$ est associée à chaque transformation élémentaire. La mesure de similarité de l'appariement inexact est ensuite obtenue en trouvant la combinaison de transformations élémentaires, associée à la fonction de coût globale minimale

$$S(\mathcal{G}_1, \mathcal{G}_2) = \min_{\mathcal{F}} \big(\sum_i C(f(\nu_i^1) = \nu_j^{2\lambda}) \big). \tag{4.57}$$

Etant donné des graphes $\mathcal{G}_1$ avec n noeuds et $\mathcal{G}_2$ avec m noeuds, $n \leq m$, si l'on effectue la recherche du minimum parmi la combinaison de transformations élémentaires possibles, la complexité calculatoire de l'appariement inexact sera de $\mathcal{O}(n^{m+1})$. Pour une recherche du minimum moins coûteuse, on aura recours à des algorithmes d'optimisation (c.f. § 4.5).
Une extension de l'appariement inexact sera employée dans la section 7.1.1 pour évaluer la similarité entre deux structures spatio-temporelles.

4.4.3 Modèles pour l'analyse de variables aléatoires multidimensionnelles

Dans cette section, on considère le cas limite des champs aléatoires de Gibbs où les variables locales X_i d'un champ aléatoire de Gibbs possèdent un voisinage d'ordre 0. On admettra donc maintenant que les n variables aléatoires $X_i \in X^n$ sont indépendantes et identiquement distribuées (i.i.d.) et qu'elles ont une dimension égale à m. Comme les variables sont i.i.d., pour éviter toutes confusions avec la dimension, la variable X_i sera notée simplement X.

Modèle normal multivarié pour l'analyse linéaire Supposons l'hypothèse d'un modèle normal multivarié pour la variable aléatoire X. En notant respectivement par M et A, le vecteur des moyennes et la matrice de covariance de la distribution gaussienne $\mathcal{N}_m$ de dimension m, la distribution de probabilités de X s'écrit

$$p(x) = \mathcal{N}_m(M, A). \tag{4.58}$$

L'hypothèse du modèle normal multivarié permettra l'analyse de structures linéaires présentes dans la distribution de probabilités multidimensionnelle de X via la décorrélation des composantes par projections linéaires[3]. C'est le principe de l'Analyse en Composantes Principales (ACP), qui sera utilisé dans la section 6.1.2.1 dans l'optique de la réduction de

[3]Dans la section 6.1.2.1, nous verrons que la décorrélation peut s'obtenir de façon équivalente via la vraisemblance de la projection.

dimensionnalité.

Modèle de mélange de gaussiennes pour l'analyse multimodale Emettons maintenant l'hypothèse d'un modèle de mélange de gaussiennes pour la variable aléatoire X. Notant par $\mathcal{N}(M_k, A_k)$ la distribution normale multivariée de vecteur de moyennes M_k et de covariance A_k, et par π_k la fonction pondératrice associée, la distribution de probabilités de X de dimension m s'écrit pour un mélange de K gaussiennes

$$p(x) \quad = \quad \sum_{k=1}^{k=K} \pi_k \mathcal{N}_m(M_k, A_k) \; , \quad \sum_{k=1}^{k=K} \pi_k = 1. \tag{4.59}$$

Le modèle de mélange de gaussiennes est capable d'approximer efficacement n'importe quelle distribution de probabilités multidimensionnelle pour laquelle aucun a priori n'est disponible. Il est particulièrement adapté pour l'analyse de distribution de probabilités multimodale significative de structures non-linéaires. Le modèle de mélange de gaussiennes sera employé dans la section 6.1.3, pour la modélisation dans l'espace multidimensionnel des caractéristiques.

Modèles non-gaussiens pour l'analyse non-linéaires L'hypothèse du modèle normal multivarié ne donne pas accès à une analyse exhaustive de la distribution de probabilités d'une variable aléatoire multivariée car elle permet seulement l'analyse de dépendances linéaires, via la décorrélation. L'hypothèse d'une distribution de probabilités non-gaussienne offre l'accès à l'analyse de structures non-linéaires, via l'indépendance statistique.
En effet, le théorème Central limite, qui est un résultat fondamental en théorie des probabilités, stipule que la distribution de la somme de variables aléatoires indépendantes tend vers une distribution gaussienne, sous certaines conditions. Donc la somme de deux variables aléatoires possède souvent une distribution de probabilités qui est plus proche d'une gaussienne que n'importe quelle des deux variables aléatoires originales. Un corollaire direct de ce théorème est que si la distribution de probabilités marginale d'une variable aléatoire est non-gaussienne (éloignée d'une distribution normale), alors elle est indépendante. Sous la contraintes que les composantes ne possèdent pas de distributions gaussiennes, le problème de la recherche de composantes indépendantes peut donc se résumer à l'élaboration et à l'optimisation (c.f. § 4.5) d'un critère performant pour l'évaluation de la non-gaussianité d'une distribution[4].
C'est la clé de l'analyse en composantes indépendantes ou encore de la poursuite de projections. Cette dernière sera employée dans la section 6.1.2.2 pour l'analyse exploratoire de structures non-gaussiennes dans l'espace multidimensionnel des caractéristiques, le but étant la réduction de dimensionnalité.

[4]Nous verrons dans la section 6.1.2.2, que cette approche est équivalente à des approches basées sur l'évaluation de l'indépendance par la vraisemblance ou l'information mutuelle.

4.5 Algorithmes d'optimisation

L'intérêt d'une méthode statistique dépend beaucoup de son applicabilité. Dans notre cas, cela signifie qu'il nous faudra trouver une stratégie qui permette de révéler les maxima d'un critère d'indépendance, ou le maximum d'une distribution de probabilités pour l'estimation de paramètres et le calcul de l'évidence de modèles, ou encore le maximum d'une fonction de coût associée à la meilleure transformation d'un appariement inexact, parmi l'explosion combinatoire des solutions possibles.
Pour répondre à ces problèmes, nous aborderons tour à tour des méthodes d'optimisation déterministes puis stochastiques.

4.5.1 Optimisations déterministes

Les techniques les plus classiques implémentent une recherche déterministe dans l'espace des solutions ou à partir de la définition de la fonction seule.

Arbres de Recherche Dans le cas de la recherche d'une solution composée d'une combinaison d'éléments, associé chacun à un coût élémentaire, l'historique du processus de recherche peut être enregistré dans une structure appelée arbre de recherche. La solution est représentée par une ramification, partant du tronc vers les branches, associé à un coût global minimum. Mais souvent la solution est difficilement accessible car il existe une explosion combinatoire de solutions ou en d'autres termes, une explosion de ramifications à évaluer. Pour faire face à ce phénomène, dans le cas d'une extension de l'appariement inexact (c.f. § 7.1.1), nous utiliserons une solution simple consistant à développer seulement les ramifications de l'arbre les plus intéressantes, tout au long de la construction de l'arbre. Cet algorithme convergera vers des minima locaux.

Méthodes de plus forte descente Pour l'évaluation du maximum d'une fonction, les plus élémentaires d'entre elles sont les méthodes de plus forte descente. Elles se basent sur une procédure itérative qui nécessite à chaque pas l'évaluation d'une direction de descente maximum (par le calcul du gradient, etc) de la fonction à maximiser. L'algorithme converge vers des maxima locaux.

Algorithme de L'Espérance-Maximisation (EM) L'algorithme EM est une méthode d'estimation de paramètres qui se place dans le contexte général de l'estimation par maximum de vraisemblance (ML). C'est essentiellement un algorithme d'optimisation itératif, qui sous certaines conditions *(Dempster, 1977)*, converge vers des valeurs de paramètres à un maximum local de la fonction de vraisemblance. L'algorithme EM présente plusieurs avantages. Entre autres, il constitue une alternative facilement implémentable à des algorithmes de plus forte descente qui peuvent nécessiter des calculs analytiques importants (e. g. calcul du gradient ou du laplacien). Par ailleurs, il est particulièrement adapté à l'estimation par maximum de vraisemblance de paramètres d'un modèle de mélange de distribu-

tions. Pour l'estimation des paramètres du modèle de mélange, l'algorithme alterne jusqu'à la convergence (1) une "étape E" évaluant les probabilités d'appartenance des points aux composantes du mélange et (2) une "étape "M" qui se base sur ces probabilités d'appartenance pour mettre à jour les estimations des paramètres du modèle.
Dans la section 6.1.3, nous nous attacherons à estimer les paramètres $\boldsymbol{\theta}_k$ et la complexité d'un modèle de mélange d'un nombre indéterminé de gaussiennes par une inférence MDL. L'estimateur ML des paramètres $\boldsymbol{\theta}_k$ du modèle de mélange de k gaussiennes sera obtenu par l'emploi de l'algorithme EM tandis que la recherche du meilleur nombre de gaussiennes (associées à leurs paramètres estimés par ML) au sens du principe MDL sera conduite par une méthode de plus forte descente.

Les techniques déterministes ont en général l'avantage de présenter un coût calculatoire restreint. Cependant, toutes les techniques déterministes présentent la grande faiblesse d'effectuer une analyse locale, et donc de pouvoir se bloquer dans des minima/maxima locaux, différents du minimum/maximum global. Pour une fonction présentant plusieurs minima/maxima, il faut donc faire très attention à l'initialisation de la procédure d'optimisation.

4.5.2 Optimisations stochastiques

L'inadaptation des optimisations déterministes, qui par leur analyse locale se bloquent souvent dans des maxima locaux, ne permet pas toujours de faire des analyses satisfaisantes. Le maximum global d'une fonction oscillante sera plus facilement révélé par les optimisations aléatoires effectuant une recherche globale. En contre partie, la découverte du maximum global nécessite un effort calculatoire généralement plus important, et parfois même prohibitif.

Face aux fonctions très oscillantes que constituent les indices utilisés en poursuite de projection, on emploie dans la section 6.1.2.2 un algorithme d'optimisation aléatoire proposé par *Huber (1990)* pour optimiser une fonction évaluant l'indépendance d'une projection bidimensionnelle d'un espace multidimensionnel. Dans l'esprit de l'algorithme du recuit simulé *(Kirkpatrick et al., 1983)*, cet algorithme intègre à la fois une recherche globale, une recherche locale dans une région de plus en plus petite autour de la solution courante et la possibilité de sortir d'un des pseudo-maxima.

4.6 Résumé

Dans ce chapitre, les points suivants ont été discutés :

- les réseaux bayésiens et les modèles hiérarchiques bayésiens pour l'apprentissage de signaux multidimensionnels ont été introduits ;

- l'estimation de paramètres et la sélection de modèles ont été dérivées des deux niveaux de l'inférence bayésienne ;

- les techniques d'extraction d'information basées sur l'entropie, telle que la discrimination minimum de Kullback, ou fondées sur la théorie du codage, telle que le principe MDL, ont été introduites ;

- les équivalences entre les méthodes entropiques et bayésiennes ont été détaillées ; la longueur de description en 2-parties MDL a notamment été présentée comme une approximation du premier ordre de l'évidence bayésienne ;

- des modèles stochastiques pour représenter des signaux spatio-temporels ont ensuite été présentés ; les champs aléatoires de Gibbs ainsi que leurs représentations sous forme de graphes ont été notamment abordées ; ils sont apparus comme des modèles génériques pour les processus stochastiques multidimensionnels ; des modèles de similarités entre graphes ont par ailleurs été détaillés ;

- les modèle gaussien et non-gaussien ont été exposés respectivement pour l'analyse de structures linéaires et non-linéaires ; le modèle de mélange de gaussiennes a été défini pour approximer efficacement toute distribution pour laquelle aucun a priori n'est disponible.

- les propriétés de méthodes d'optimisation déterministes et stochastiques, souvent nécessaires dans le processus de modélisation, ont été exposées ; en particulier, ont été définies : les arbres de recherches pour lutter contre les problèmes d'explosion combinatoire, l'algorithme EM pour le cas particulier de l'estimation par ML des paramètres d'un modèle de mélange, ainsi qu'un algorithme d'optimisation aléatoire pour révéler le maximum global d'une fonction oscillante.

Les techniques présentées dans ce chapitre sont appliquées à la modélisation hiérarchique de structures spatio-temporelles qui sont présentées dans le chapitre suivant.

Chapitre 5

Concept de fouille visuelle d'information dans les séries temporelles d'images

Dans ce chapitre, on présente un concept de fouille visuelle d'information destiné à la compréhension de scènes dynamiques. En accord avec la définition de la section 3.3.2, la compréhension de scènes dynamiques a pour objectif l'identification de structures spatio-temporelles précises et la découverte de règles régissant ces structures.
La compréhension de scène dynamique est relative à un utilisateur désireux d'exploiter un certain jeu de données. Pour atteindre ces objectifs, la compréhension de scènes dynamiques devra donc employer un processus d'apprentissage qui sera conduit par l'intervention d'un utilisateur, et contraint par les données. Dans cette optique, le concept développé dans cette thèse se base sur un apprentissage non-supervisé puis supervisé des probabilités d'un modèle hiérarchique bayésien. Ce modèle hiérarchique bayésien formalise les liens entre les structures spatio-temporelles des STIS et les différentes sémantiques d'un utilisateur. Il concrétise donc un canal de communication entre les données et l'utilisateur au travers duquel l'information est représentée de façon hiérarchique. En accord avec le concept décrit dans la section 3.2.3, la hiérarchie se décompose en deux sous-parties : une partie représentation "signal" issue d'un apprentissage non-supervisé, et une partie représentation "sémantique" issue d'un apprentissage supervisé.
Dans la première partie de ce chapitre, on présentera d'abord le modèle hiérarchique bayésien formalisant les liens probabilistes entre les données et les sémantiques d'utilisateurs. On décrira ensuite, de façon synthétique, comment transite l'information au travers des différents niveaux de la hiérarchie. Finalement, on introduira les modèles et les processus d'inférence relatifs aux différents niveaux de la hiérarchie et, on fournira par la même les références aux chapitres détaillant ces traitements. Dans la deuxième partie de ce chapitre, les choix de représentation visuelle pour l'apprentissage interactif seront introduits puis justifiés.

5.1 Modélisation hiérarchique bayésienne de l'information

Dans le domaine de la fouille d'information, les modèles hiérarchiques bayésiens sont des modèles puissants (c.f. § 3.3.2). Leur puissance dépend de la quantité d'information perdue à chaque inférence réalisée au niveau supérieur[1].

5.1.1 Hiérarchie pour la compréhension de scènes dynamiques

Pour étendre la fouille d'information à la compréhension de scènes dynamiques, un modèle hiérarchique bayésien à 6 niveaux est utilisé pour lier la source d'information D, qui représente les structures spatio-temporelles des STIS, aux différentes sémantiques $\boldsymbol{\mathcal{A}_\nu}$ des utilisateurs. La hiérarchie est définie par la relation entre les variables aléatoires

$$D \rightarrow \Theta \rightarrow \Psi \rightarrow C \rightarrow \mathcal{G}_k \rightarrow \mathcal{A}_\nu, \tag{5.1}$$

qui sont associées à des niveaux de représentation de l'information des STIS de plus en plus raffinés. $\boldsymbol{D}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\boldsymbol{C}$, $\boldsymbol{\mathcal{G}_k}$ et $\boldsymbol{\mathcal{A}_\nu}$ sont des réalisations de ces variables aléatoires. La signification de chacune d'elles sera explicitée dans la section suivante. Nous avons vu au chapitre précédent dans la présentation des modèles hiérarchiques bayésien que l'inférence d'un niveau supérieur dans la hiérarchie dépend du niveau directement inférieur et, conditionnellement à ce dernier, est indépendante des niveaux indirectement inférieurs. On obtient donc la probabilité conjointe associée au modèle hiérarchique exprimée par

$$p(\boldsymbol{\mathcal{A}_\nu} \mid \boldsymbol{\mathcal{G}_k})p(\boldsymbol{\mathcal{G}_k} \mid \boldsymbol{C})p(\boldsymbol{C} \mid \boldsymbol{\psi})p(\boldsymbol{\psi} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{D})p(\boldsymbol{D}). \tag{5.2}$$

Chaque passage à un niveau supérieur de la hiérarchie constitue un apprentissage. Pour effectuer les inférences de chaque niveau supérieur de la hiérarchie, on maximise ces probabilités ou, par la règle de Bayes, celles du modèle hiérarchique inverse

$$p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{\psi})p(\boldsymbol{\psi} \mid \boldsymbol{C})p(\boldsymbol{C} \mid \boldsymbol{\mathcal{G}_k})p(\boldsymbol{\mathcal{G}_k} \mid \boldsymbol{\mathcal{A}_\nu})p(\boldsymbol{\mathcal{A}_\nu}). \tag{5.3}$$

Pour définir les distributions de probabilités conditionnelles ci-dessus, les inférences bayésiennes ou basées sur des mesures entropiques, seront appliquées sur des familles de modèles stochastiques adaptées aux signaux multidimensionnels.

[1]Le système de fouille d'images I2M en particulier (c.f. annexe A.1) est constitué par un modèle hiérarchique bayésien contenant des représentations de l'information possédant des niveaux croissants d'abstraction sémantique. Des recherches menées ces dernières années ont permis d'évaluer la qualité de ce modèle *(Datcu et Seidel, 2002)*. Ce modèle hiérarchique a influencé celui développé dans cette thèse. Cette influence s'exprime à plusieurs niveaux de la hiérarchie, qui seront explicités par la suite.

On distingue deux sous-parties dans ce modèle hiérarchique :

- un apprentissage non-supervisé $D \rightarrow \mathcal{G}_k$: d'abord à partir d'une STIS $\boldsymbol{D}$, on fait l'inférence automatique de graphes $\boldsymbol{\mathcal{G}_k}$ qui modélisent les trajectoires de clusters dynamiques et qui codent les structures spatio-temporelles de la STIS ; cet apprentissage non-supervisé est explicité dans le chapitre 6 ;

- un apprentissage interactif $\mathcal{G}_k \rightarrow \mathcal{A}_\nu$: ensuite on lie interactivement les sous-graphes $\boldsymbol{\mathcal{G}_k}$ de la collection, représentant une structures spatio-temporelle de la STIS, à une collection de sémantiques $\boldsymbol{\mathcal{A}_\nu}$ d'utilisateurs ; on engendre ainsi un étiquetage sémantique des structures spatio-temporelles des STIS ; cet apprentissage supervisé est explicité dans le chapitre 7.

L'inférence du graphe représente la partie coûteuse du traitement mais permet de coder automatiquement et efficacement le contenu informatif d'une STIS. Par contre, l'apprentissage des sémantiques attachées aux motifs du graphe présente une faible complexité calculatoire. A partir de quelques exemples positifs et négatifs fournis par un utilisateur, le système apprend interactivement des sémantiques complexes.

Présentons de façon plus précise, et dans l'ordre dicté par la hiérarchie, ce que représente chacune de ces variables aléatoires puis par quels mécanismes d'inférence chacune de ces variables aléatoires est obtenue.

5.1.2 Organisation des différents espaces et variables de la hiérarchie

La figure 5.1 schématise et détaille l'organisation des différents niveaux de la hiérarchie. On se référera à la section 2.4 pour les notations des espaces et des formes de représentation des signaux.

$\boldsymbol{D}$: les données. Le niveau le plus bas représente les données $\boldsymbol{D}$, c'est à dire les structures spatio-temporelles des STIS. Chaque structure spatio-temporelle de la STIS est identifiée par une certaine fenêtre temporelle ainsi que par un certain masque spatial. C'est dans l'espace spatio-temporel $\mathcal{E}_{r,t,\theta}$ que les données sont représentées. Celles-ci constituent un signal de la forme $x(r,t,\theta)$.

$\boldsymbol{\theta}$: les caractéristiques. Pour passer au niveau supérieur, on procède à l'extraction d'une collection de caractéristiques $\boldsymbol{\theta}$. Cette collection est constituée par les d ensembles de caractéristiques $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{t_1}, ..., \boldsymbol{\theta}_{t_d}\}$ extraites au niveau des n pixels, aux différents instants t_i. On obtient ainsi une description du signal spatialement et temporellement localisée.
On utilise diverses caractéristiques telles que les caractéristiques spectrales, de texture et de forme. On fait l'hypothèse simplificatrice que ces caractéristiques ne sont pas dépendantes, et l'on analyse celles-ci de façon indépendante.

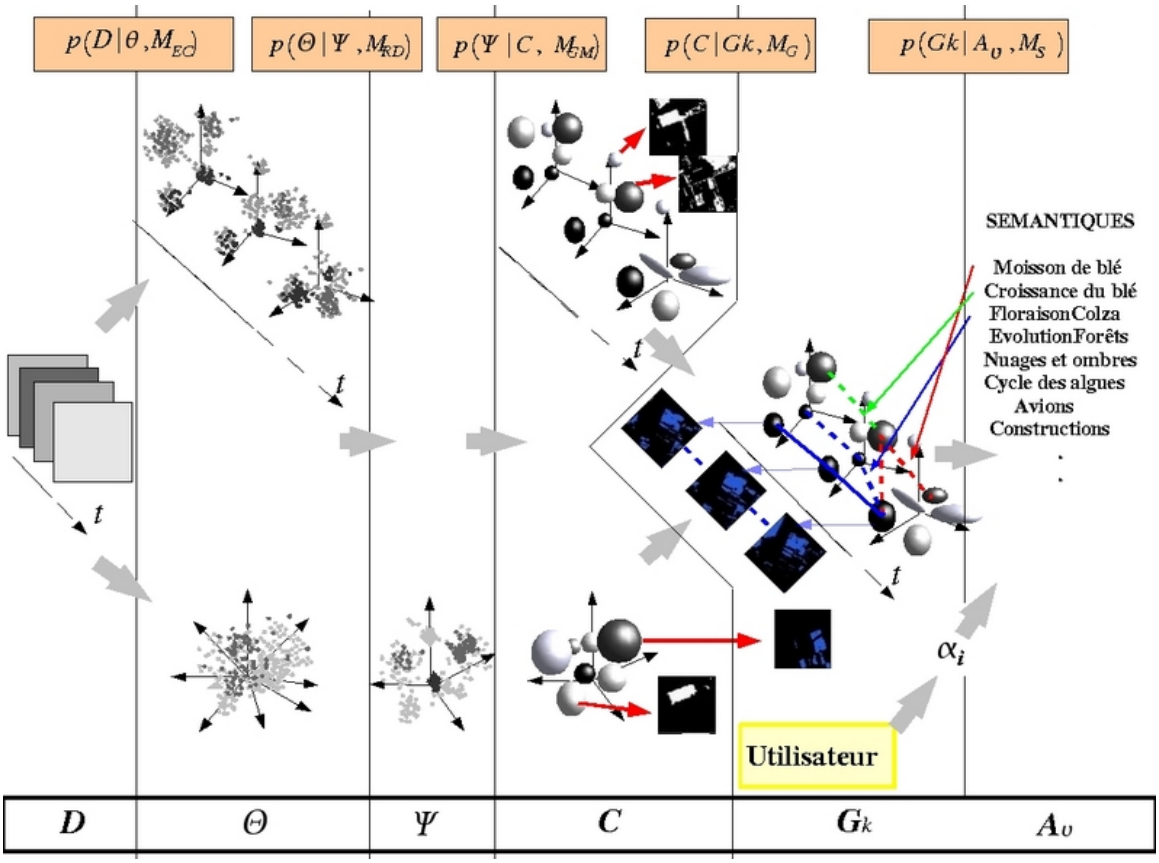


FIG. 5.1 – *Schéma de la modélisation hiérarchique bayésienne du contenu informatif des STIS. Cette hiérarchie permet de relier les structures spatio-temporelles contenues dans les données* $\mathbf{D}$*, à des sémantiques d'utilisateur* $\mathcal{A}_\nu$*. On distingue deux sous-parties dans cette modélisation hiérarchique : un apprentissage non-supervisé du signal qui permet de coder les données* $\mathbf{D}$*, c'est à dire les structures spatio-temporelles des STIS, par des graphes* $\mathcal{G}_k$*, associés à des classifications dynamiques ; et un apprentissage supervisé interactif qui permet un étiquetage sémantique de ces graphes. Des modèles stochastiques puissants et des inférences bayésiennes ou entropiques sont employés pour déduire chaque nouveau niveau de la hiérarchie.*
Les différentes étapes d'apprentissage sont les suivantes. Des caractéristiques sont d'abord extraites de chaque image composant la STIS. Ces caractéristiques sont placées dans l'espace MT des caractéristiques (en bas) et dans les espaces TL des caractéristiques (en haut). Ensuite, pour contourner la "malédiction de la dimensionnalité", des projections "intéressantes" sont extraites de l'espace MT des caractéristiques. Puis, par l'emploi d'un modèle de mélange, on partitionne les différents espaces afin de constituer des collections $\boldsymbol{C}$ *de clusters et de classes. Finalement, pour achever l'apprentissage non-supervisé, on modélise l'espace des caractéristiques évolutives, en inférant des graphes de trajectoires de clusters MT (en bas), constitués par des clusters TL (en haut). Parallèlement, dans l'espace spatio-temporel, on associe à chaque trajectoire de cluster MT (e. g. trajectoire bleu), une classification dynamique (e. g. classifications bleues). Un étiquetage sémantique interactif des graphes est ensuite effectué par l'intervention d'un utilisateur.*

Les caractéristiques $\boldsymbol{\theta}$ sont alors représentées, comme l'illustre la figure 5.1, dans deux types d'espaces de caractéristiques différents :

- les espaces TL des caractéristiques $\{\mathcal{E}_{\theta_{t_i}};\ i \in [1,d]\}$, qui sont les espaces où les caractéristiques relatives à chaque image de la STIS constituent des signaux de la forme $\{x_r(\theta_{t_i}); i \in [1,d]\}$,
- et l'espace MT des caractéristiques $\mathcal{E}_{\theta_{t_1},...,\theta_{t_d}}$, qui est l'espace formé par tous les sous-espaces TL, et où les caractéristiques constituent un signal de la forme $x_r(\theta_{t_1},...,\theta_{t_d})$.

La complémentarité des représentations, qui a déjà été abordée dans la section 2.4, sera rediscutée dans la section 6.1.1.2.

$\boldsymbol{\psi}$: les caractéristiques projetées. L'espace MT des caractéristiques peut se révéler être un espace de dimension élevée de par sa formation : l'espace formé à partir des composantes de tous les espaces TL. C'est ici que se situe une grande difficulté connue sous la dénomination de "malédiction de la dimensionnalité" : pour un nombre fixé d'échantillons, plus le nombre de dimensions de l'espace de représentation est grand, moins cet espace est peuplé. Un des corollaires bien identifié de ce phénomène est la difficulté d'inférences dans des espaces de dimension trop élevée. Toutefois, l'information contenue dans ces espaces multidimensionnels peut souvent se condenser et être représentée dans un espace avec moins de dimensions. Par exemple, dans les STIS, il existe une grande redondance d'information entre les bandes spectrales entre les différents échantillons temporels. On met donc en oeuvre des techniques de réduction de dimensionnalité pour projeter, en utilisant l'opérateur de projection $\mathcal{P}(.)$, les caractéristiques MT dans un nouvel espace MT réduit de caractéristiques, qui est noté $\mathcal{E}_{\mathcal{P}(\theta_1,...,\theta_d)}$, où les caractéristiques constituent un signal de la forme $x_r(\mathcal{P}(\theta_1,...,\theta_d))$. On obtient ainsi n nouvelles caractéristiques MT, qui forment avec les d ensembles de n caractéristiques TL, un ensemble de caractéristiques que l'on notera $\boldsymbol{\psi}$.

$\boldsymbol{C}$: les partitions en clusters et classes. Pour apprendre comment sont distribuées les caractéristiques $\boldsymbol{\psi}$ et ainsi réduire leurs descriptions, on modélise les distributions de probabilités des caractéristiques dans les espaces TL et dans l'espace MT réduit. En employant un modèle multimodal composé par un mélange de distributions, on discrimine une collection de partitions $\boldsymbol{C}$ pour les différentes espaces de caractéristiques. Chaque partition est constituée par un ensemble d'éléments, et chacun de ces éléments représente une distribution du mélange. Chacun de ces éléments, appelé cluster, caractérise une sous-population de caractéristiques dont la description se réduit aux attributs associés à une distribution du mélange. Le processus de groupage des sous-populations se nomme clustering. Pour les espaces TL de caractéristiques, on engendre donc par modélisation puis clustering, d collections de K_{t_i} clusters $\{C^1_{t_i},...,C^{K_{t_i}}_{t_i}\}$ localisées aux différents temps t_i avec $i \in [1,d]$. Ces clusters sont qualifiés de clusters TL. En exécutant la même procédure avec l'espace MT des caractéristiques, on obtient une collection de K_{MT} clusters $\{C^1_{MT},...,C^{K_{MT}}_{MT}\}$ appelés clusters MT.
Parallèlement au clustering dans les espaces TL de caractéristiques, la modélisation produit aussi des partitions dans l'espace spatial $\mathcal{E}_r$. On représente la STIS dans cet espace sous la forme de signaux $\{x_{\theta_{t_i}}(r);\ i \in [1,d]\}$, appelés classifications TL. Chaque élément

$C^j_{t_i}(r)$ d'une classification TL est appelé classe TL et a un cluster TL associé $C^j_{t_i}$. De façon analogue, parallèlement au clustering dans l'espace MT réduit de caractéristiques, une partition est produite dans l'espace spatial $\mathcal{E}_r$. On représente la STIS dans cet espace sous la forme du signal $x_{\theta_{t_1},\ldots,\theta_{t_d}}(r)$, appelée classification MT. Elle est composée d'éléments $C^j_{MT}(r)$, appelés classes MT, qui sont associées un à un avec les clusters MT C^j_{MT}.

$\mathcal{G}_k$: les graphes et les classifications dynamiques. On cherche maintenant à partir des clusters et des classes TL et MT, à modéliser la STIS représentée dans l'espace des caractéristiques dynamiques $\mathcal{E}_{t,\theta}$ et dans l'espace spatio-temporel $\mathcal{E}_{r,t}$. La première représentation $x_r(t,\theta)$ exhibe l'information sur l'évolution de la distribution des caractéristiques, tandis que la deuxième représentation exhibe l'information sur l'évolution spatiale des structures. Ces représentations contiennent une information complémentaire et décomposent en deux la représentation de la STIS dans son espace naturel. Comme remarqué dans la section 2.4, les espaces TL de caractéristiques sont la décomposition de l'espace des caractéristiques évolutives. Par l'introduction de liens de causalité entre les clusters TL et entre les classes TL, on formalise les trajectoires de clusters et de classes dynamiques. L'information de causalité est apportée par l'introduction de la représentation MT. La complémentarité des informations fournies par les deux représentations sera par la suite détaillée (c.f. § 6.1.1.2). Par un processus d'inférence détaillé par la suite, on obtient 2 types de caractérisations des structures spatio-temporelles : un graphe de trajectoires de clusters dynamiques associées à une collection de classifications dynamiques.
Le graphe de trajectoires forme une collection de graphes $\mathcal{G}_k$, où chacun de ces graphes représente une structure spatio-temporelle de la STIS.

$\mathcal{A}_\nu$: les sémantiques d'utilisateurs. Les collections de graphes et de classifications dynamiques, obtenues par l'apprentissage non-supervisé décrit précédemment, forment une description synthétique et objective des structures spatio-temporelles de la STIS. Finalement, basé sur cette description objective, on relie l'intérêt d'un utilisateur pour certaines sémantiques $\mathcal{A}_\nu$ aux structures spatio-temporelles possédant ces sémantiques. On s'applique donc maintenant à attacher aux collections de graphes et de classifications dynamiques des labels sémantiques. Cet apprentissage est effectué interactivement via la communication au système d'exemples positifs et négatifs de graphes et de classifications dynamiques, introduits par un utilisateur.

5.1.3 Modèles et inférences des différents niveaux hiérarchiques

Extraction de caractéristiques (EC) : $D \rightarrow \theta$. Nous avons vu que la première étape d'inférence est l'extraction de caractéristiques au niveau de chaque pixel de chaque image de la STIS. On emploie ici des techniques issues du système I2M présenté dans l'annexe A.1 et dont l'efficacité a déjà été évaluée *(Datcu et Seidel, 2002)*. Les modèles et les processus d'inférence sont décrits brièvement.
On applique des modèles stochastiques M_{EC} pour capturer les structures spatiales des images. La qualité de ces modèles paramétriques par rapport aux données $\boldsymbol{D}$ s'évalue par

la vraisemblance

$$p(\boldsymbol{D} \mid \boldsymbol{\theta}, \mathcal{M}_{EC}),$$

qui associe une probabilité pour un certain vecteur de paramètres $\boldsymbol{\theta}$. Le processus d'extraction d'information se fait par une estimation bayésienne du vecteur de paramètres. Des caractéristiques d'images, dépendant des modèles employés, sont ainsi obtenues.
Un exemple de modèle $\mathcal{M}_{EC}$ est la famille des champs aléatoires de Gibbs pour l'extraction de texture ou simplement les réflectances des images multispectrales pour les caractéristiques spectrales. Evidement, ce dernier cas ne nécessite pas de modèle probabiliste.
Le processus d'extraction de caractéristiques sera plus amplement détaillé dans la section 6.1.1.1.

Réduction de Dimensionnalité (RD) : $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}$. La modélisation d'espaces de dimension élevée est un problème complexe qui peut souvent être résolu par l'emploi au préalable de techniques de réduction de dimensionnalité.
Ces techniques projettent un espace des caractéristiques de dimension m de façon à retrouver un nombre $l \leq m$ de composantes jugées intéressantes. Ces méthodes font l'hypothèse que les caractéristiques sont distribuées selon un certain modèle stochastique $\mathcal{M}_{RD}$. Les caractéristiques projetées $\boldsymbol{\psi}$ sont donc créées par une application injective, souvent constituée par une matrice de projection linéaire. La qualité de la projection peut s'évaluer par la vraisemblance

$$p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathcal{M}_{RD}).$$

Comme aucun a priori n'est incorporé dans le processus d'inférence, l'estimateur du maximum de vraisemblance peut donc s'employer pour trouver la meilleure matrice de projection pour un certain jeu de caractéristiques.
Mais en pratique le plus souvent, l est égale ou est légèrement inférieur à m. Afin d'obtenir une réduction de dimensionnalité, on garde seulement les projections qui possèdent un indice de pertinence au-dessus d'un certain seuil fonction d'un certain critère.
L'état de l'art a montré qu'une méthode très utilisée est l'Analyse en Composantes Principales (ACP). Dans cette analyse, le modèle est une distribution Normale multivariée. On cherche à rendre la matrice de covariance diagonale, c'est à dire à décorréler les composantes. Cette analyse de l'espace des caractéristiques permet la capture de structures linéaires. Comme on le verra par la suite, la décorrélation équivaut à la maximisation de la vraisemblance de la projection. La décorrélation s'obtient à partir d'un processus de diagonalisation qui présente un faible coût calculatoire. Pour obtenir une réduction de dimensionnalité par l'ACP, on conserve seulement l'ensemble des projections associées aux plus grandes valeurs propres qui préservent une certaine proportion de l'énergie du signal. Le détail de l'ACP sera présenté dans la section 6.1.2.1.
Dans notre cas, de par les multiples sources constituant chaque image, l'hypothèse d'une distribution Normale multivariée n'est généralement pas valide. Néanmoins, l'ACP permet de sélectionner, avec un faible coût calculatoire, un sous-espace contenant une proportion importante de l'énergie du signal, dans un espace de dimension trop élevée. Les structures non-linéaires présentes dans le sous-espace extrait sont ensuite analysées par la technique de poursuite de projection (PP). Celle-ci fait l'hypothèse d'un modèle $\mathcal{M}_{RD}$

multivarié non-gaussien. On capture ainsi des projections indépendantes présentant des structures non-linéaires. Comme on le verra par la suite, dans ce cas, la maximisation de la non-gaussianité est équivalente à la maximisation de la vraisemblance de la projection. L'évaluation de l'indépendance s'obtient à partir d'un processus itératif employant une optimisation stochastique. Pour obtenir une réduction de dimensionnalité par PP, on conserve seulement l'ensemble des projections estimées comme les composantes indépendantes les plus structurées. Le détail de l'analyse par PP sera présenté dans la section 6.1.2.2.

Clustering et classification : $\boldsymbol{\psi \rightarrow C}$. Afin de décrire synthétiquement l'espace des caractéristiques, on procède au clustering de cet espace, qui engendre parallèlement une classification dans l'espace image. Pour capturer les différentes structures que forment les caractéristiques dans cet espace, on effectue un groupage des caractéristiques par similarité en fonction d'un modèle de mélange. Le modèle de mélange de gaussiennes $\mathcal{M}_{GM}$ est bien adapté à l'estimation de distribution multimodale. Par ailleurs, ce modèle permet une approximation efficace d'une distribution pour laquelle aucun a priori n'est disponible. La qualité d'un certain modèle de mélange de gaussiennes engendrant les clusters $\boldsymbol{C}$, s'évalue pour un certain jeu de caractéristiques projetées $\boldsymbol{\psi}$ par la vraisemblance

$$p(\boldsymbol{\Psi} \mid \boldsymbol{C}, \mathcal{M}_{GM}).$$

Le modèle de mélange de gaussiennes est une famille de modèles paramétriques. Elle possède une complexité variable dépendant du nombre de gaussiennes K. Pour la modélisation, il faut donc sélectionner le nombre de gaussiennes du modèle et estimer les matrices de covariances, les vecteurs moyennes et les pondérations associées. L'inférence du mélange de gaussiennes est effectuée par le principe MDL qui emploie une longueur de description en 2-étapes. Notons que précédemment, il a été remarqué que cette sélection de modèles est une approximation du premier ordre de l'évidence bayésienne. L'algorithme EM et une méthode de plus forte descente sont utilisés pour trouver le minimum du critère. L'inférence MDL du mélange de gaussiennes sera plus longuement développée dans la section 6.1.3.

Inférence de graphes caractérisant les structures spatio-temporelles : $\boldsymbol{C \rightarrow \mathcal{G}_k}$ Pour modéliser la dynamique des structures spatio-temporelle, on formalise les trajectoires des clusters MT au travers des clusters TL. De plus, on introduit une information spatiale dans le processus de modélisation par la prise en compte des classes TL et MT. Pour appliquer cette idée, on emploie le modèle de trajectoire de clusters $\mathcal{M}_T$. Ce modèle est une distribution de probabilité sur la collection de clusters $\boldsymbol{C}$ conditionnée par un certain graphe. La qualité du modèle de trajectoire de clusters pour les graphes $\boldsymbol{\mathcal{G}_k}$ inclus dans le graphe $\mathcal{G}$, est évaluée en fonction d'une certaine collection $\boldsymbol{C}$ de clusters et de classes TL et MT par la vraisemblance

$$p(\boldsymbol{C} \mid \boldsymbol{\mathcal{G}_k}, \mathcal{M}_T). \tag{5.4}$$

N'ayant pas d'a priori sur le graphe, l'inférence se fait par maximum de vraisemblance. Nous verrons que la maximisation de la vraisemblance équivaut à maximiser une fonction inversement proportionnelle à la divergence de Kullback-Leibler. Par cette équivalence,

l'inférence se résume au calcul analytique de divergences de Kullback-Leibler, et se fait avec un coût calculatoire réduit. Cette inférence sera explicitée dans la section 6.2.1.
Le processus de caractérisation des trajectoires, c'est à dire la caractérisation du graphe et des classifications dynamiques est alors appliquée. Il engendre une collection d'attributs qui sont associés aux noeuds et aux arcs des graphes. Le processus de caractérisation de trajectoires sera présenté plus en détails dans la section 6.2.2.

Modélisation de la sémantique de l'utilisateur : $\mathcal{G}_k \rightarrow \mathcal{A}_\nu$ L'apprentissage non-supervisé constitué par les niveaux d'inférences inférieurs a engendré les graphes $\mathcal{G}_k$ caractérisant les structures spatio-temporelles. En se basant sur cette caractérisation objective de l'information de la STIS, on s'attache maintenant à associer des sémantiques $\mathcal{A}_\nu$ à cette collection de graphes.
Pour cela, on introduit d'abord un modèle paramétrique de similarité $S_\phi(\mathcal{G}_0, \mathcal{G}_k)$ qui évalue la similarité entre deux graphes $\mathcal{G}_0$ et $\mathcal{G}_k$. Les paramètres sont constitués par un graphe de référence $\mathcal{G}_0$ et un vecteur de paramètres ϕ qui pondère la contribution de chaque type d'attributs de graphe dans la mesure de similarité. Il sera présenté dans la section 7.1.1.
On estime les paramètres de ce modèle pour parvenir à une représentation sémantique. En effet, on admet qu'un jeu de paramètres particulier correspond à une similarité particulière, qui elle-même correspond à une certaine sémantique. Donc, pour un vecteur de paramètre particulier ϕ et un certain graphe de référence $\mathcal{G}_0$, alors plus un graphe $\mathcal{G}_k$ est similaire à $\mathcal{G}_0$, plus la probabilité qu'il possède la sémantique de l'utilisateur $\mathcal{A}_\nu$ sera grande. A partir de cette idée, on définit le modèle sémantique $\mathcal{M}_S$ qui est une distribution de probabilités sur les graphes $\mathcal{G}_k$, conditionnée par les sémantiques $\mathcal{A}_\nu$. Ce modèle sémantique est paramétrique car il dépend de ϕ et de $\mathcal{G}_0$. Pour chaque sémantique $\mathcal{A}_\nu$ et pour chaque graphe $\mathcal{G}_k$, on peut pour un certain modèle paramétrique $\mathcal{M}_S$, calculer les probabilités de vraisemblance

$$p(\mathcal{G}_k \mid \mathcal{A}_\nu, \mathcal{M}_S).$$

Pour estimer les paramètres du modèle $\mathcal{M}_S$, un apprentissage interactif est mis en oeuvre. Celui -ci se base sur un modèle de Dirichlet, introduit dans la section 7.1.2.2, qui permet la mise à jour des probabilités d'un réseau bayésien. Cette mise à jour s'effectue via les hyper-paramètres $\{\alpha_i\}$ du modèle de Dirichlet, interactivement estimés par la communication au système d'exemples de graphes possédant une sémantique positive $\mathcal{A}_\nu$ et négative $\neg\mathcal{A}_\nu$. Des exemples de sémantique positive et négative désignent respectivement des exemples d'intérêt ou de non-intérêt pour l'utilisateur. Cet apprentissage interactif sera explicité dans la section 7.1.2.2
L'étiquetage de la sémantique $\mathcal{A}_\nu$ est obtenu par un seuillage des probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k, M)$ évaluées pour chaque graphe de la collection. Après chaque nouvel apprentissage, les probabilités a posteriori des graphes sont réévaluées. Le modèle sémantique ainsi que l'étiquetage sémantique seront présentés dans la section 7.1.

5.2 Supports visuels pour l'apprentissage interactif

Par ailleurs, on propose des représentations visuelles de l'information qui aident le processus d'apprentissage interactif.
Différentes informations visuelles sont proposées. Ces informations visuelles sont accessibles grâce à différentes représentations, que l'on appellera supports visuels. Comme nous l'avons souligné dans la section sur la fouille visuelle d'information, l'utilisateur peut visualiser deux types d'information : des informations accessibles par des représentations du signal visuellement pertinentes et des informations liées à des représentations sémantiques.
Les représentations du signal, c'est à dire ici les représentations des structures spatio-temporelles des STIS, ont la vocation de décomposer l'information complexe de données multidimensionnelles pour l'adapter à la perception humaine. Par exemple, on cherchera pour les STIS des représentations exhibant plus clairement l'information sur l'évolution spatiale des structures et l'information sur l'évolution des distributions des caractéristiques relatives à ces structures.

Nous employons les représentations de STIS suivantes :

- des séries temporelles d'images dont la dynamique des couleurs a été adaptée pour plus d'homogénéité visuelle entre les différents échantillons temporels ; plus précisément, on effectue une adaptation dynamique des histogrammes des images pour que celles-ci suivent la même loi Normale ; ces objets sont représentés dans l'espace $\mathcal{E}_{r,t,\boldsymbol{\theta}}$;
- des graphes de trajectoires, représentatifs de l'évolution des distributions de caractéristiques relatives aux différentes structures spatio-temporelles ; ces graphes sont représentés dans l'espace des caractéristiques évolutives $\mathcal{E}_{t,\boldsymbol{\theta}}$;
- des classifications dynamiques qui sont représentatives de l'évolution spatiale des structures ; elles sont représentées dans l'espace spatio-temporel $\mathcal{E}_{r,t}$;

L'utilisateur pourra donc explorer la STIS au travers de ces différentes représentations. Chaque exemple positif ou négatif fourni par l'utilisateur durant l'apprentissage pourra être défini au travers de l'une de ces représentations. Pour cela, il désignera dans ces représentations des fenêtres spatio-temporelles contenant les structures d'intérêt.
Parallèlement à la visualisation des données, pour supporter l'apprentissage interactif, on est intéressé par la visualisation de représentations sémantiques de l'information. Pour un certain degré d'apprentissage sémantique, l'utilisateur visualise l'étiquetage sémantique courant et des indicateurs de pertinence.

Dans la suite, nous introduisons et justifions, les différents supports visuels choisis dans cette thèse pour la fouille visuelle d'information dans les STIS. La section 8.1.2 détaillera les interactions Homme-Machine s'effectuant au travers de ces représentations.

Visualisation de séries temporelles d'images normalisées. La STIS représentée dans son espace naturel constitue une représentation de référence pour l'utilisateur qui peut accéder à l'information dans sa globalité. Cependant, afin d'améliorer l'homogénéité

visuelle des caractéristiques de réflectances entre les différents échantillons temporels, on procède à l'adaptation non-linéaire des histogrammes. Cette transformation se propose donc de limiter visuellement, le problème de discontinuité temporelle des radiométries soulevée dans la partie 2.3 : les incertitudes liées aux modèles physiques employés pour calculer les réflectances engendrent des distorsions radiométriques qui sont ensuite corrigées par un ajustement linéaire. Mais, il subsiste toujours des hétérogénéités visuelles entre les réflectances des images de la STIS. Ces hétérogénéités ne sont pour la plupart pas naturelles car elles apparaissent entre des images séparées par seulement quelques jours d'intervalle.
Pour atteindre une meilleure homogénéité visuelle entre images de la série, on fait l'hypothèse que pour chaque bande spectrale de chaque image de la STIS, les réflectances recherchées possèdent la même distribution normale. Par adaptation dynamique, des STIS aux réflectances normalisées sont crées. Cette nouvelle STIS sera visuellement plus intuitive, même si l'homogénéité crée est artificielle. Par ailleurs, d'un point de vue plus pratique, on utilise les formats standards de visualisation (e. g. JPEG ou MPEG-4) qui utilisent un codage sur 8 bits. Les transformations non-linéaires d'histogrammes des réflectances, codées sur 10 bits et non-uniformément réparties, permettent une meilleure quantification que celle obtenue avec des transformations linéaires. Ces modifications d'histogramme seront explicitées dans la section 8.1.2.

Visualisation d'un graphe multidimensionnel. Les graphes de trajectoires de clusters dynamiques présentés plus haut constituent des représentations permettant la visualisation des dépendances entre structures spatio-temporelles.
Ces graphes sont représentés dans l'espace des caractéristiques évolutives et révèlent des informations non accessibles dans l'espace naturel de représentation des STIS : des évolutions de distributions multidimensionnelles relatives aux structures spatio-temporelles. L'utilisateur visualise les différentes composantes marginales de ce graphe multidimensionnel, c'est à dire les évolutions des paramètres des clusters, leurs apparitions ou disparitions, ou encore leurs scissions ou leurs fusions.

Visualisation de classifications dynamiques. L'information sur l'évolution spatiale fait défaut dans la représentation graphique précédente. L'utilisateur accède à celle-ci par la visualisation des classifications dynamiques dans l'espace spatio-temporel. Ces représentations révèlent en particulier des informations sur la fusion ou la scission de classes.

Visualisation d'étiquetage sémantique et d'indicateurs de pertinence Les structures spatio-temporelles associées à des probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ supérieures à 0.5 sont marquées en rouge. La couleur rouge de ces labels est nuancée proportionnellement aux probabilités a posteriori.
Afin de qualifier les degrés de contribution de chaque type d'attributs de graphe dans la définition d'une sémantique particulière, on mesure pour chaque type d'attributs de graphe le degré de distinction entre les sémantiques positives et négatives. Pour concrétiser cette idée proposée par *Schroder et al. (2000)*, on dérive une mesure de divergence de Kullback-Leibler symétrisée pour chaque élément ϕ_i du vecteur $\phi = \{\phi_1, \phi_2, ...\}$ qui paramétrise le

modèle de similarité. Chaque divergence est exprimée par

$$D_{sym}\big(p(\phi_i \mid \mathcal{A}_\nu),\, p(\phi_i \mid \neg\mathcal{A}_\nu)\big). \tag{5.5}$$

Parce que les différents éléments $\{\Phi_i\}$ sont supposés indépendants entre eux, on peut directement dériver le degré global de l'apprentissage sémantique par le produit des mesures de divergence. Le calcul de ces indicateurs de pertinence sera détaillé dans la section 7.2. Les différentes mesures de divergence sont visualisées sous forme de barres d'histogramme, avec des hauteurs, fonctions de ces mesures.
Complémentairement aux mesures de divergence, d'autres indicateurs de pertinence sont dérivés : la convergence du processus d'apprentissage est évaluée et l'homogénéité des structures pour une certaine sémantique est mesurée.
Après chaque itération du processus d'apprentissage, les différentes mesures de pertinence sont mises à jour.

5.3 Résumé

Dans ce chapitre, nous avons introduit un concept de fouille visuelle d'information pour la compréhension de scènes dynamiques des STIS. Les points suivants caractérisent l'approche adoptée :

- le concept se base sur une modélisation hiérarchique bayésienne des STIS qui permet de relier les structures spatio-temporelles constituant la STIS à des sémantiques d'utilisateurs ;

- la hiérarchie se compose de deux parties principales : l'apprentissage non-supervisé et l'apprentissage interactif des structures spatio-temporelles ; la première partie nécessite un important coût calculatoire tandis que la deuxième partie possède une complexité calculatoire réduite ;

- la phase d'apprentissage non-supervisé est concrétisée par l'inférence de graphes et de classifications dynamiques, qui modélisent respectivement l'évolution temporelle des caractéristiques du signal et l'évolution spatiale du signal ; des modèles stochastiques puissants ainsi que des inférences bayésiennes ou des inférences basées sur des mesures entropiques sont employées pour apprendre les distributions des variables des différents niveaux de la hiérarchie ;

- la phase d'apprentissage supervisé emploie un modèle sémantique paramétrique qui se base sur une mesure de similarité de graphes ; Un certain vecteur de paramètres du modèle sémantique est supposé correspondre à une sémantique particulière ; Un modèle de Dirichlet appliqué à un réseau bayésien permet l'inférence interactive de ce vecteur de paramètres particulier. ; l'apprentissage est effectué par l'intermédiaire d'un utilisateur qui fournit au système des exemples de structures spatio-temporelles

de sémantique négative et positive ;

- l'interactivité du système se base sur une interface Homme-Machine au travers de laquelle l'utilisateur explore différentes représentations du signal. Par ailleurs l'utilisateur visualise des indicateurs de pertinence qualifiant le degré d'apprentissage sémantique.

Chapitre 6

Inférence d'un graphe de trajectoires de structures spatio-temporelles

Dans ce chapitre, on s'attache à détailler l'apprentissage non-supervisé du concept introduit dans le chapitre 5. Les inférences des différents niveaux de la hiérarchie relatifs à cet apprentissage non-supervisé seront séquentiellement développées.

6.1 Modélisation de différentes représentations du signal

On présente les deux premières phases d'apprentissage qui se composent de l'extraction des caractéristiques, suivie de leur analyse.

6.1.1 Extraction et représentation des caractéristiques du signal

L'extraction de caractéristiques de la STIS constitue le premier niveau d'inférence du modèle hiérarchique. Elle est effectuée indépendamment pour chaque image de la STIS. Les caractéristiques extraites seront ensuite placées conjointement dans l'espace MT des caractéristiques et dans l'espace TL des caractéristiques. Le choix de ces espaces de représentation permet une décomposition de l'information. Mais avant d'aborder ce point, discutons des modèles et des inférences adoptées pour l'extraction de caractéristiques.

6.1.1.1 Extraction de caractéristiques des images

On applique des modèles stochastiques M_{EC} pour capturer les structures spatiales et spectrales des images. Etant donné une certaine image perçue comme une réalisation $\boldsymbol{D}$ d'un processus stochastique, pour une valeur particulière du vecteur de paramètre $\boldsymbol{\theta}$, on évalue la qualité du modèle en utilisant la vraisemblance $p(\boldsymbol{D} \mid \boldsymbol{\theta}, \mathcal{M}_{EC})$. L'évaluation de cette distribution permet l'estimation des paramètres $\widehat{\boldsymbol{\theta}}$ du modèle $\mathcal{M}_{EC}$. Les estimations de paramètres sont finalement utilisées pour constituer les vecteurs de caractéristiques.

FIG. 6.1 – *Caractéristiques de réflectance proche infra-rouge, rouge et vert-jaune, extraites des 4 premières images de la STIS ADAM (du 15/10/00 au 09/11/00) visualisées en RVB.*

L'extraction de caractéristiques spectrales est directe : chaque pixel de l'image est associé à un vecteur de caractéristiques $\boldsymbol{\theta}$ composé par les 3 différentes valeurs de réflectance. La figure 6.1 présente les 4 premières images de réflectance extraites dans une fenêtre spatiale de 200x200 pixels de la STIS ADAM. Cette image est visualisée en utilisant un espace de couleur RVB, ou le rouge, le vert et le bleu représentent respectivement la bande spectrale proche infra-rouge, rouge et vert-jaune.

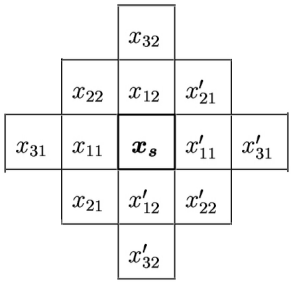


FIG. 6.2 – *Voisinage $\partial x_s = \{(x_{ij}, x'_{ij})\}$ du troisième ordre pour un pixel central x_s.*

De même, pour l'extraction de caractéristiques de texture, on associe à chaque pixel de l'image bidimensionnelle un vecteur de caractéristiques. Celui-ci est issu d'un processus de modélisation. La modélisation se base sur les deux niveaux d'inférence bayésienne menant à l'estimation des paramètres du modèle et au calcul de l'évidence du modèle. La famille de modèles stochastiques utilisée est constituée par les champs aléatoires de Gibbs (c.f. § 4.4.1). Cette approche a déjà été utilisée et évaluée par *Schroeder et al. (1998)* pour l'extraction d'information dans les images satellites. On rappelle l'approche adoptée. La modélisation fait l'hypothèse que l'image possède la propriété de Markov à un certain ordre : la statistique du niveau de gris x_s d'un pixel dans l'image dépend seulement des niveaux de gris ∂x_s des pixels appartenant à un certain voisinage de taille limitée autour de x_s. Un voisinage du troisième ordre qui est illustré dans la figure 6.2 est choisi d'une manière empirique. La probabilité du niveau de gris d'un pixel x_s est donnée en accord

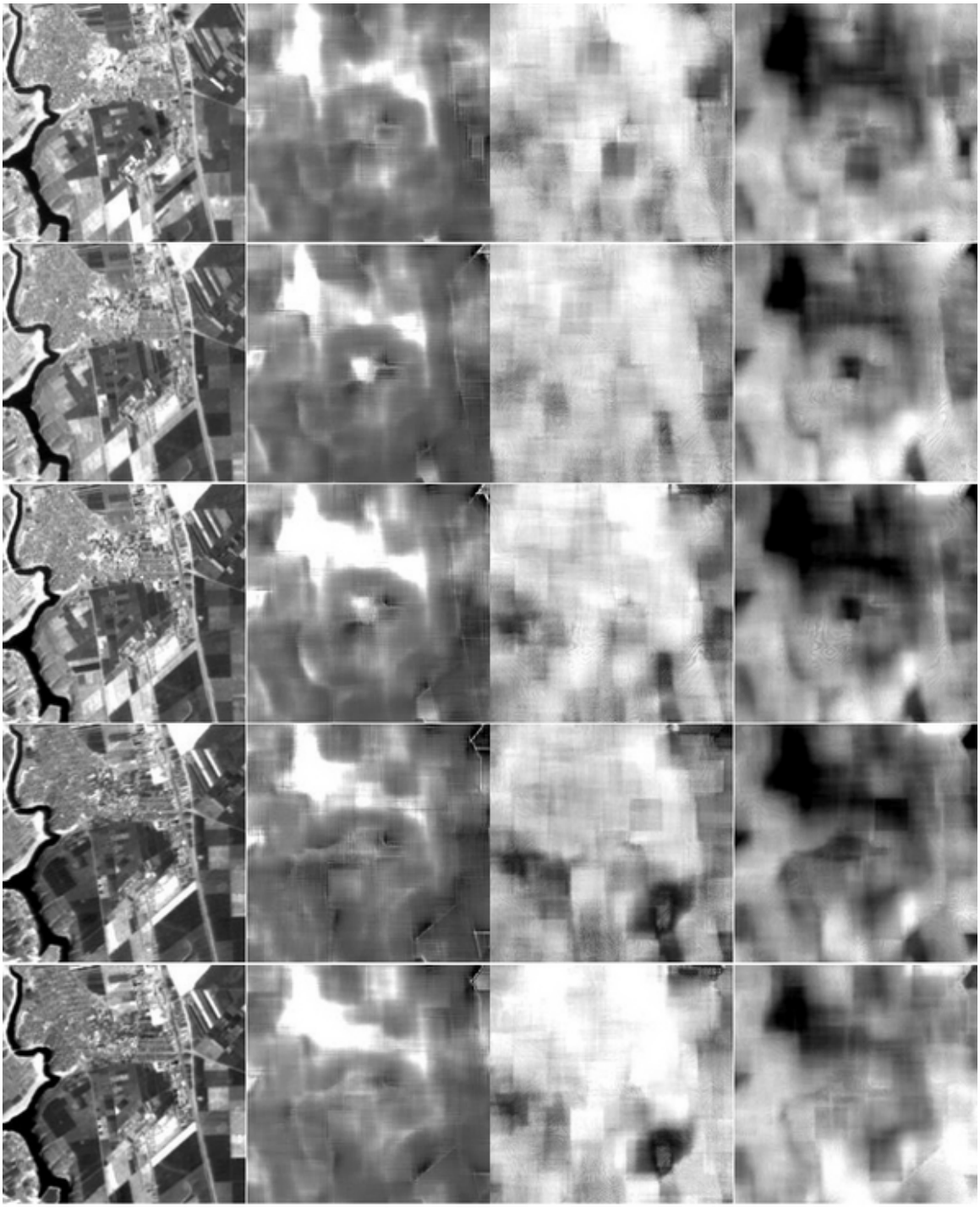


FIG. 6.3 – *Caractéristiques de texture extraites des 5 premières images de la STIS ADAM (du 15/10/00 au 14/11/00). Première colonne $\boldsymbol{D}$: bande spectrale proche infra-rouge ; deuxième colonne $\widehat{\theta_0}$: paramètre d'ordre zéro ; troisième colonne $|\widehat{\boldsymbol{\theta}}|$: norme du vecteur de paramètre. Quatrième colonne $p(D \mid \mathcal{M}_{EC})$: évidence du modèle.*

avec l'Eq. 4.56 par

$$p(x_s \mid \partial x_s, \boldsymbol{\theta}) = \frac{1}{Z_s} \exp(-U(x_s \mid \partial x_s, \boldsymbol{\theta})), \tag{6.1}$$

où Z_s joue le rôle du facteur de normalisation donné par la somme de tous les niveaux de gris possibles x_s du pixel central. $U(x_s \mid \partial x_s, \boldsymbol{\theta})$ est une fonction d'énergie qui dépend du vecteur de paramètres $\boldsymbol{\theta}$. On utilise la fonction d'énergie du modèle auto-binomial qui est définie par l'expression

$$U(x_s \mid \partial x_s, \boldsymbol{\theta}) \;=\; -log\binom{G}{x_s} - x_s\eta \tag{6.2}$$

où G représente le niveau de gris maximum et $\binom{k}{n}$ représente un coefficient binomial. La quantité

$$\eta \;=\; \theta_0 + \sum_{i,j} \theta_{ij} \frac{x_{ij} + x_{ij}^{'}}{G} \tag{6.3}$$

reflète l'influence des niveaux de gris des pixels voisins avec le vecteur de paramètres $\boldsymbol{\theta}$. Ce dernier comprend le paramètre θ_0 d'ordre 0 représentant une sorte d'auto-interaction, et les paramètres $\{\theta_{ij}\}$, où chaque paramètre θ_{ij} décrit l'interaction entre x_s et la paire $(x_{ij}, x_{ij}^{'})$ du voisinage.
L'estimation du vecteur de paramètres est effectuée par l'estimateur du minimum de l'erreur quadratique conditionnelle (CLS). Pour le détail du calcul de l'évidence du modèle $p(D \mid \mathcal{M}_{EC})$, on réfère le lecteur à *Schroeder (1998)*. Pour chaque pixel central, on constitue un vecteur de caractéristiques (et méta-caractéristiques) composé par le paramètre estimé d'ordre zéro $\widehat{\theta_0}$, la norme $|\widehat{\boldsymbol{\theta}}|$ du vecteur de paramètres estimés et l'évidence du modèle $p(D \mid \mathcal{M}_{EC})$.
L'extraction des caractéristiques de texture est effectuée à partir de la bande spectrale proche infra-rouge des images car, pour les zones rurales, elle présente la plus grande dynamique. Pour estimer les paramètres et l'évidence du modèle de texture du 3-ième ordre (c.f Fig. 6.2) associé à chaque pixel de l'image, on emploie une fenêtre d'estimation de taille de 31x31 pixels fixée empiriquement. La figure 6.3 présente la norme, le paramètre d'ordre zéro et l'évidence du modèle de texture estimée dans une fenêtre spatiale de 200x200 pixels de la bande spectrale proche infra-rouge des 5 premières images de la STIS ADAM. Ces résultats montrent qu'il existe une certaine stabilité temporelle des caractéristiques de textures dans les zones urbaines et pour certaines autres structures. Des changements de texture sont observés dans les zones rurales de l'image lors de la maturation des champs à droite de l'image.

6.1.1.2 Représentation dans un espace multitemporel et des espaces temporellement localisés

Dans cette section, on réfère le lecteur aux notations, définitions et remarques sur les espaces de représentations des STIS introduites dans la section 2.4. L'extraction de caractéristiques est effectuée au niveau de chaque pixel des différentes images de la STIS.

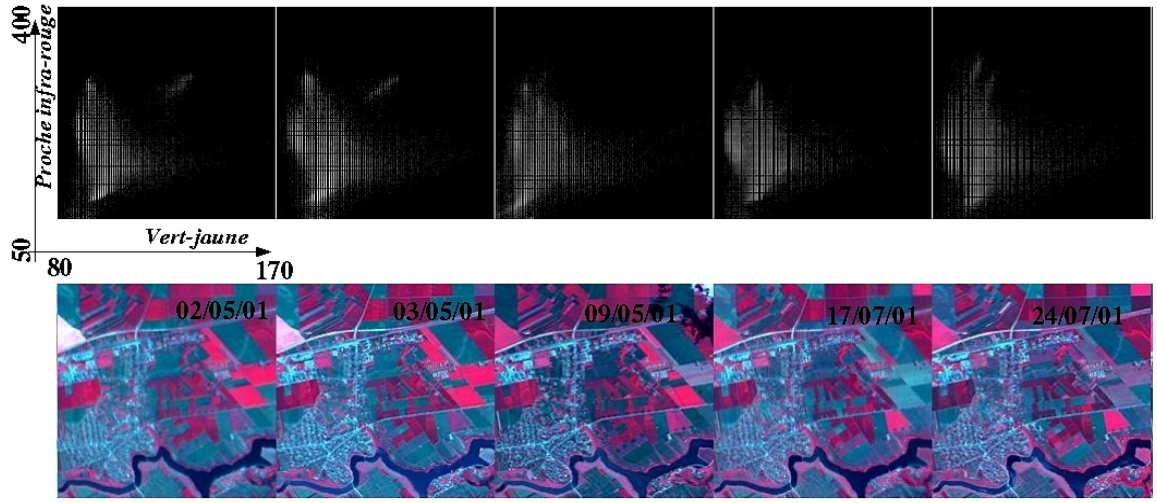


FIG. 6.4 – *Représentation de la série d'images dans 5 espaces TL de caractéristiques consécutifs (haut), associées à 5 images de la STIS (bas).*

Pour ces d différentes images, représentons dans les d espaces TL des caractéristiques correspondants $\{\mathcal{E}_{\theta_{t_1}}, ..., \mathcal{E}_{\theta_{t_d}}\}$ les différentes collections extraites de vecteurs de caractéristiques $\{\boldsymbol{\theta}_{t_i};\ i \in [1, d]\}$. Dans chacun de ces espaces TL, on représente la STIS sous la forme du signal $\{x_r(\theta_{t_i}); i \in d\}$, qui contient une information sur les caractéristiques à un temps donné, et indirectement par l'index spatial r, une information spatiale TL. La figure 6.4 présente 5 espaces TL consécutifs, pour des caractéristiques spectrales.
Parallèlement, on représente les d collections de vecteurs de caractéristiques dans un même espace : l'espace MT des caractéristiques $\mathcal{E}_{\theta_{t_1},...,\theta_{t_d}}$. On y représente le signal $x_r(\theta_{t_1}, ..., \theta_{t_d})$, qui contient l'information sur l'évolution temporelle des caractéristiques et, indirectement par l'index spatial r, l'information spatiale globale des structures spatio-temporelles

L'espace MT contient l'information spatio-temporelle globale sur la STIS y compris l'information TL. Les structures y sont bien caractérisées du fait de la dimension élevée. La séparabilité des structures dans cet espace sera donc plus importante que dans les espaces TL. Mais en contrepartie, il est bien connu que l'analyse des espaces de dimension élevée comme celui-ci posent des problèmes. Nous verrons dans la section suivante des techniques pour lutter contre la malédiction de la dimensionnalité. Toutefois, parce que ces techniques possèdent des limites d'analyse, nous emploierons conjointement les espaces TL pour une extraction d'information plus détaillée.

6.1.2 Réduction de dimensionnalité

L'espace MT des caractéristiques est un espace dont la dimension est élevée. Celle-ci est égale au nombre d'échantillons temporels fois la dimension de l'espace TL des

caractéristiques. L'espace MT contient l'information sur l'évolution de la distribution des caractéristiques dans le temps. On analyse les caractéristiques dans cet espace afin d'en extraire l'information pertinente. Pour cela on considère que la collection de caractéristiques $\boldsymbol{\theta}^n$ sont les n réalisations d'une variable aléatoire $\Theta = \{\Theta_1, ..., \Theta_m\}$ de dimension m dont on cherche à modéliser la distribution de probabilités. Une modélisation directe de la distribution des caractéristiques multidimensionnelles est dans ce cas sérieusement restreinte par la limitation appelée la malédiction de la dimensionnalité *(Duda et al., 2000)*. Néanmoins, l'information contenue dans un espace multidimensionnel peut souvent être représentée dans un espace de dimension moins élevée.

Les techniques d'analyse multidimensionnelle étudient des méthodes statistiques pour projeter la collection des n caractéristiques $\boldsymbol{\theta}^n$ de dimension m de façon à retrouver $l \leq m$ composantes. On fait l'hypothèse que les caractéristiques sont distribuées selon un certain modèle stochastique $\mathcal{M}_{RD}$. Parmi ces techniques, on se restreint dans ces travaux de thèse à celles qui effectuent des projections linéaires de l'espace à analyser. On admet que les n caractéristiques $\boldsymbol{\theta}^n$ de dimension m sont le résultat de l'application d'une matrice de mélange $A = \{a_{ij}\}$ sur les n caractéristiques sources $\boldsymbol{\psi}^n = \{\psi^1, ..., \psi^n\}$ de dimension l . En notant par $\Psi = \{\Psi_1, ..., \Psi_l\}$ la variable aléatoire de dimension l engendrant les réalisations $\boldsymbol{\psi}^n$, le modèle de mélange linéaire s'exprime par

$$\Theta_i = a_{i1}\Psi_1 + a_{i2}\Psi_2 + ... + a_{il}\Psi_l; \ \forall i \tag{6.4}$$

ou encore sous forme matricielle, par $\Theta = A\Psi$. Pour retrouver la matrice de mélange, on maximise la vraisemblance de la projection $p(\boldsymbol{\theta}^n \mid \boldsymbol{\psi}^n, \mathcal{M}_{RD})$. Suivant le modèle $\mathcal{M}_{RD}$ recherché, on substituera cette estimation par des approches équivalentes.
En se basant sur ce modèle de transformation linéaire, on expose dans la suite l'analyse en composantes principales (ACP) et l'analyse par poursuite de projection (PP). Ces analyses présentent des équivalences avec l'analyse discriminante *(Benzecri ,1973)*. L'ACP est adaptée pour l'analyse linéaire alors que l'analyse par PP permet l'exploitation des structures non-linéaires et multimodales et autorise une réduction de dimensionnalité supplémentaire. La distribution des caractéristiques dans l'espace des projections obtenues est ensuite modélisée par un modèle de mélange (c.f. § 6.1.3), permettant ainsi la discrimination des différentes structures.

6.1.2.1 Analyse en composantes principales

L'ACP fait l'hypothèse que les caractéristiques possèdent une distribution Normale multivariée. L'analyse cherche une combinaison linéaire de composantes marginales possédant une distribution de probabilités Normale multivariée $\mathcal{M}_{RD} = \mathcal{N}_m(0, A_{diag})$ de matrice de covariance diagonale A_{diag}. La décorrélation des composantes est donc recherchée. Pour obtenir la décorrélation, il est plus simple d'effectuer une transformation de Karhunen-Loeve des caractéristiques ou de façon équivalente, de maximiser l'inertie de la projection, que de maximiser la vraisemblance $p(\boldsymbol{\theta}^n \mid \boldsymbol{\psi}^n, \mathcal{M}_{RD})$. Maximiser l'inertie de la projection revient à minimiser l'erreur quadratique moyenne entre le k^{eme} vecteur réalisation θ^k et sa

projection orthogonale dans la base recherchée. L'inertie de la projection est définie par

$$J = \frac{1}{n} \sum_{k=1}^{n} |\psi^k|^2_{\mathcal{N}} \tag{6.5}$$

où $\mathcal{N}$ souligne le fait que les caractéristiques ont été au préalable centrées et réduites. Pour obtenir une réduction de dimensionnalité, on emploie l'énergie moyenne du signal donnée par

$$\mathcal{W} = \sum_{i=1}^{m} \sigma_i^2 = \sum_{i=1}^{m} \lambda_i \tag{6.6}$$

òu σ_i^2 désigne la variance des distributions marginale Ψ_i, et λ_i désigne les valeurs propres de la matrice d'autocorrélation des caractéristiques $\boldsymbol{\theta}^n$ *(Spataru, 1987)*. On conserve seulement l'ensemble des $q \leq m$ premières composantes principales, associées aux plus grandes variances, qui ensemble restituent un certain pourcentage α de l'énergie du signal $\mathcal{W}$. De manière équivalente, on peut calculer l'énergie restituée en utilisant les plus grandes valeurs propres de la matrice d'autocorrélation. Le pourcentage de l'énergie restituée par la sélection des q premières composantes est donné par

$$\alpha = \frac{\sum_{i=1}^{q} \sigma_i^2}{\mathcal{W}} * 100. \tag{6.7}$$

Une réduction de dimensionnalité basée sur l'ACP permet la capture de structures linéaires présentant une forte variance. De plus, si l'hypothèse de gaussiannité est valide, les composantes extraites seront indépendantes. Par ailleurs, l'ACP possède une faible complexité calculatoire, car elle est simplement basée sur la recherche des valeurs et vecteurs propres d'une matrice d'autocorrélation.
Mais dans notre cas, les caractéristiques spectrales et texturales à analyser sont issues d'un mélange de plusieurs sources (e.g. parcelle de différentes cultures, villes, rivière, etc). L'hypothèse d'une distribution Normale multivariée n'est donc généralement pas valide car les distributions de probabilités des projections sont multimodales. En conséquences, les composantes extraites présenteront des structures non-linéaires. Les projections seront décorrélées mais pas indépendantes, et ordonnées selon un critère de variance. Il est possible que dans ce cas, on omette des composantes de faible variance présentant des structures non-linéaire. Néanmoins, l'ACP sera ici utile pour pré-selectionner, avec une complexité calculatoire réduite, un ensemble de q projections de l'espace MT, tout en préservant une proportion importante de l'énergie du signal. Une analyse par PP (section suivante) permettra ensuite l'extraction de $l \leq q$ composantes présentant des structures non-linéaires, dans le sous espace formé par les composantes pré-selectionnées par l'ACP.

On effectue l'ACP de la STIS représentée dans l'espace MT des caractéristiques dans une fenêtre spatiale de 800x800 pixels. Pour les 38 images, 114 dimensions composent l'espace MT des caractéristiques spectrales. La figure 6.5 présente les 2 premières composantes principales représentées dans l'espace image et dans l'espace des caractéristiques.

Pour les caractéristiques spectrales, on observe dans l'espace image un fort contraste pour les deux 2 premières composantes principales qui correspond à des grandes variances dans

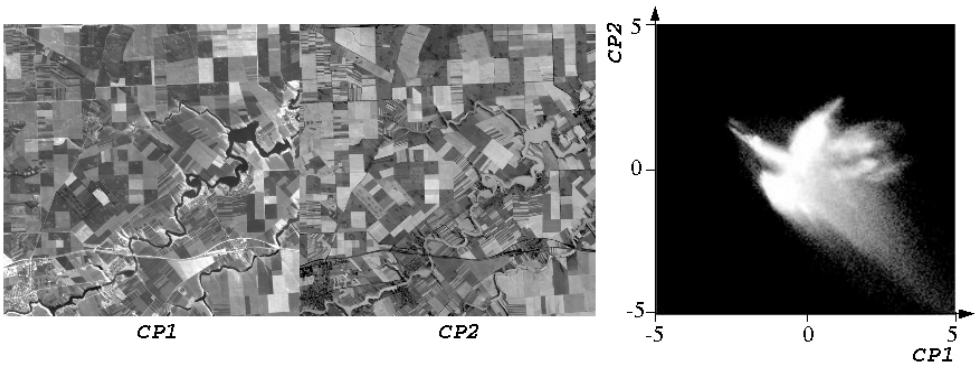


FIG. 6.5 – *2 premières composantes principales spectrales, représentées dans l'espace image et des caractéristiques.*

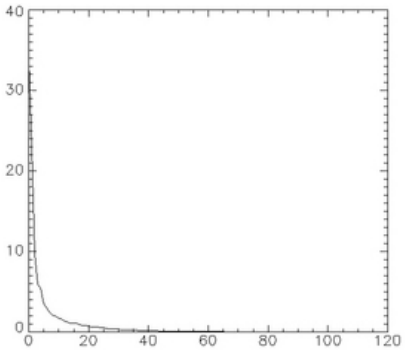


FIG. 6.6 – *Répartition de l'énergie sur les composantes principales pour les caractéristiques spectrale. L'axe des abscisses supporte les 114 composantes principales rangées en fonction de leurs variances et l'axe des ordonnées présente l'énergie relative à chaque composante.*

l'espace des caractéristiques. En accord avec les études sur le comportement de l'ACP en détection de changements, les premières composantes ont tendance à regrouper des structures spatio-temporelles qui demeurent assez stables dans le temps. On note aussi une décorrélation entre les composantes au niveau des parcelles agricoles ou encore des rivières. Mais, on n'obtient pas une réelle indépendance entre les composantes. On observe dans l'espace des caractéristiques qu'il existe bien des structures non-linéaires de forte variance. Suite à cette analyse, on procède à la sélection des 20 premières composantes principales qui représentent 90 pourcent de l'énergie du signal. Dans ce cas, la réduction de dimensionnalité est importante car l'énergie est concentrée sur les premières composantes principales. La figure 6.6 illustre la répartition de l'énergie du signal sur les composantes principales.

L'ACP est utilisée pour la sélection d'un sous espace d'analyse dans l'espace MT. On réduit ainsi la complexité d'analyses postérieures tout en préservant une importante proportion de l'énergie du signal. Mais, l'ACP ne permet pas une réduction de dimensionnalité importante car elle se limite à l'analyse de structures linéaires et ne prend pas en considération la nature non-linéaire des structures dans l'espace MT des caractéristiques.

6.1.2.2 Analyse exploratoire par Poursuite de Projection

Pour effectuer une réduction de dimensionnalité de l'espace MT par l'extraction de projection présentant des structures non-linéaires intéressantes, on emploie les techniques exploratoires de PP. Ces méthodes explorent l'espace multidimensionnel à analyser et extraient des projections linéaires qui sont les maxima d'une fonction, appelée indice de projection, qui évalue un critère de non-gaussianité.
L'analyse par PP représente une solution adaptée à la nature non-linéaire des structures d'une STIS. Une réduction de la dimensionnalité de l'espace MT des caractéristiques est accomplie par l'extraction des projections indépendantes qui présentent le plus de structures. Dans la suite, on présentera d'abord les liens entre l'analyse par PP et les autres analyses. On exposera ensuite brièvement les équivalences des méthodes d'estimation de non-gaussianité. Puis, on introduira un indice de projection basé sur la distance du chi-deux*(Huber, 1985)* qui présente de bonnes propriétés. Pour la recherche des maxima de l'indice de projection, on introduira un algorithme d'optimisation stochastique basé sur un algorithme développé par *Posse (1993)*. On montrera ensuite comment l'on parvient à une réduction de dimensionnalité. Finalement, on présentera des résultats sur la réduction de dimensionnalité supplémentaire obtenue par une analyse par PP de l'espace formé à partir des composantes principales sélectionnées.

Equivalences de l'analyse par PP avec l'ACP et l'Analyse en Composantes Indépendantes (ACI). L'analyse par PP est en un sens une généralisation de L'ACP. En effet, l'ACP peut être perçue comme une adaptation de l'analyse par PP où des projections linéaires sont extraites en fonction d'un critère de variance maximum, avec la contrainte que ces projections soient décorrélées.

Par ailleurs, l'analyse par PP et l'ACI sont des analyses équivalentes sous certaines conditions. En effet, l'ACI cherche un ensemble de composantes indépendantes entre elles ne possédant pas une distribution gaussienne, et l'analyse par PP cherche des projections qui maximisent un critère de non-gaussianité. Or, le critère de non-gaussianité est équivalant au critère d'indépendance sous la contrainte que les composantes indépendantes ne soient pas gaussiennes (c.f. § 4.4.3). En d'autres termes, si l'hypothèse de non-gaussianité est valide, les composantes estimées par ACI et les projections extraites par PP sont les mêmes et celles-ci sont indépendantes entre-elles. Intéressons nous maintenant au cas de l'espace MT des caractéristiques. Lorsque nous avons discuté de l'ACP, dans cet espace, les distributions de probabilités des projections recherchées sont multimodales et donc non-gaussiennes. En conséquence, l'analyse par PP sera équivalente à l'ACI, et les projections extraites seront indépendantes.
Contrairement à l'ACI, l'analyse par PP permet de quantifier le niveau d'indépendance de chaque projection et donc d'ordonner les composantes. En sélectionnant les composantes les plus structurées ou encore les plus indépendantes, on pourra obtenir une réduction de dimensionnalité par PP.

Estimation de la non-gaussianité. Pour estimer les $l \leq q$ projections indépendantes $\Psi = \{\Psi_1, ..., \Psi_l\}$ à partir des $\Theta = \{\Psi_1, ..., \Psi_q\}$, plusieurs méthodes équivalentes existent. Basées sur les équivalences clarifiées dans le paragraphe précédent entre la non-gaussianité et l'indépendance, des méthodes d'estimation de projections indépendantes, maximisent des mesures de non-gaussianité. Ces mesures approximent la divergence de Kullback-Leibler entre des distribution projetées et la distribution gaussienne. Elles se basent donc sur la maximisation

$$\max_{\Psi} D(p_\Psi, p_\mathcal{N}) \tag{6.8}$$

où $p_\mathcal{N}$ est la distribution Normale de dimension l, centrée, réduite et aux composantes décorrélées, et où p_Ψ est la distribution de la variable recherchée Ψ. On remarque que l'hypothèse d'indépendance entre les composantes permet de décomposer l'expression de $D(p_\Psi, p_\mathcal{N})$ en une somme de divergences mono-dimensionnelles, ou bi-dimensionnelles, ou de dimension plus élevée. Basés sur cette décomposition de la divergence, l'analyse par PP et certains algorithmes d'ACI effectuent donc la recherche des maxima d'un indice de projection mono-dimensionnelle, bi-dimensionnelle, etc.
On peut montrer que la minimisation de l'information mutuelle entre les différentes projections est une approche équivalente à la maximisation de la divergence de Kullback-Leibler. La maximisation de la vraisemblance de la projection $p(\boldsymbol{\theta}^n \mid \boldsymbol{\psi}^n, \mathcal{M}_{RD})$, où $\mathcal{M}_{RD}$ est simplement une distribution non-gaussienne, est aussi une approche théoriquement équivalente. Toutefois, il existe en pratique une légère différence car pour l'estimation, le modèle $\mathcal{M}_{RD}$ doit être connu. Dans certains cas où l'information a priori sur la nature des composantes indépendantes est fausse, cette approche peut donc engendrer des résultats incorrects. Pour plus de détails sur les équivalences des approches, on réfère le lecteur à *Lee et al. (2000)*.

Indice de projection. Plus la dimension des projections est élevée, plus riche est l'information contenue dans celles-ci. Par exemple, certaines structures décelables dans le plan ne

peuvent pas l'être dans des projections mono-dimensionnelles. On sera donc tenté de choisir un indice de dimension la plus élevée possible. Mais, pour des raisons de simplicité, on recherche les maxima d'un indice bidimensionnel afin d'approximer la maximisation de la divergence de Kullback-Leibler de l'Eq. 6.8. Plus précisément, comme l'on recherche des composantes-plans indépendantes, la divergence de Kullback-Leibler se décompose en une somme bidimensionnelle

$$D(p_{\Psi}, \mathcal{N}_{2D}) = \sum_{i=2,\, i\, pair}^{l} D(p_{(\Psi_{i-1},\Psi_i)}, p_{\mathcal{N}_{2D}}) \tag{6.9}$$

où $p_{\mathcal{N}_{2D}}$ est la distribution Normale bivariée, centrée, réduite et aux composantes décorrélées et où $p_{(\Psi_{i-1},\Psi_i)}$ est la distribution conjointe de (Ψ_{i-1}, Ψ_i). Mais l'estimation de ces quantités nécessite un important coût calculatoire. On approxime donc chaque élément de cette somme, par un Indice de Projection (IP) bi-dimenionnelle :

$$IP(\Psi_{i-1}, \Psi_i) \simeq D(p_{(\Psi_{i-1},\Psi_i)}, p_{\mathcal{N}_{2D}}). \tag{6.10}$$

Donc si l'approximation est efficace, l'ensemble de projections obtenues par le maximum de la divergence $D(\Psi, \Psi_{gauss})$ est identique à l'ensemble de projections bidimensionnelles correspondant aux maxima de l'indice de projection $IP(\Psi_{i-1}, \Psi_i)$.
Pour approximer la divergence, on utilise un indice de projection basée sur la distance du chi-deux. Cet indice est orthogonalement invariant, sensible aux structures dans le centre de la distribution et robuste aux queues lourdes. Mais son principal avantage est qu'il présente une faible complexité calculatoire. On note par (α, β) le plan formé par les deux composantes recherchées (Ψ_{i-1}, Ψ_i) et, comme le propose *Posse (1993)*, on partitionne ce plan en 48 boîtes $\{B_k\}$. On définit alors l'indice de projection du chi-deux par

$$IP(\alpha, \beta) = \frac{4}{\pi} \int_0^{\frac{\pi}{4}} PI_{\alpha,\beta}(\eta) d\eta, \tag{6.11}$$

où

$$IP_{\alpha,\beta}(\eta) = \sum_{k=1}^{48} \frac{\Big(\int\int_{B_k} \{dF_{\alpha,\beta}(\eta) - d\mathcal{N}_{2D}\} \Big)^2}{\int\int_{B_k} d\mathcal{N}_{2D}}. \tag{6.12}$$

$dF_{\alpha,\beta}(\eta)$ et $d\mathcal{N}_{2D}$ désignent respectivement la distribution théorique des données et la distribution gaussienne $\mathcal{N}_{2D}$ dans les boîtes. Pour l'estimation, les $dF_{\alpha,\beta}(\eta)$ théoriques sont remplacées par les distributions empiriques $d\widehat{F}_{\alpha,\beta}(\eta)$.

Optimisation stochastique Afin de trouver chaque plan (α, β) correspondant à un couple de composantes indépendantes (Ψ_{i-1}, Ψ_i), il nous faut rechercher le maximum global de l'indice de projection. On emploie une optimisation stochastique proposée par *Huber (1985)*. Cet algorithme intègre à la fois une recherche globale, une recherche locale dans une région de plus en plus petite autour de la solution courante et la possibilité de s'éloigner d'un des pseudo-maxima. La gestion de cet algorithme est très flexible, la profondeur de la recherche dépendant essentiellement de trois paramètres : le nombre de

départs aléatoires, la taille initiale du voisinage autour de la solution courante et le temps de recherche entre deux diminutions de la taille des voisinages visités. De plus, la précision de la recherche est aisément contrôlée par un critère d'arrêt spécifié par l'utilisateur. Toutes les opérations responsables de cette gestion étant très simples, le temps de calcul propre à celle-ci est négligeable, contrairement à l'algorithme du recuit simulé qui nécessite un temps de calcul supplémentaire. Autrement dit, s'agissant d'un algorithme hautement itératif, sa rapidité se mesure directement sur le temps d'évaluation de la fonction à maximiser. Pour plus de détails théoriques et sur l'implémentation on réfère le lecteur à la thèse de *Posse (1993)*.
Mais, après avoir obtenu un maximum global, pour trouver les autres composantes indépendantes correspondant au autres maxima et éviter de retrouver le maximum global, on procède à la déstructuration de la projection. Cette idée a initialement été proposée par *Friedman (1987)*. Le principe consiste à transformer la structure non-gaussienne trouvé lors de la maximisation de l'indice de projection, par une distribution Normale bivariée, centrée réduite et aux composantes décorrélées. Une alternative intéressante à la déstructuration est de forcer la décorrélation entre la projection recherchée et les composantes indépendantes déjà trouvées *Lee et al (2000)*. Ainsi, on relance l'optimisation jusqu'à ce que tous les plans structurés de l'espace soient révélés.

Réduction de dimensionnalité. Mais, en pratique et particulièrement dans l'espace MT des caractéristiques, on ne rencontre jamais une distribution gaussienne parfaite, et à moins d'extraire un nombre de composantes indépendantes égale au nombre de dimensions de l'espace d'origine, on se doit d'estimer le nombre de composantes indépendantes. *Posse (1990)* dérive pour l'index du chi-deux une approximation à un terme des P-Valeurs. Une P-Valeur représente la probabilité d'erreur, c'est à dire la probabilité qu'une composante-plan extraite soit du bruit. Une valeur critique de l'indice de projection est associée à une P-Valeur. En particulier, en utilisant des approximations, *Posse* montre qu'une P-Valeur de 0.01 correspond à une valeur critique de l'indice de projection égale à

$$\frac{63.55190 + 12.34898.q}{n}. \tag{6.13}$$

où n représente le nombre de réalisations et q la dimension de l'espace à analyser. Par l'emploi de cette approximation, on obtiendra ainsi $l \leq q$ composantes indépendantes.

Exemple d'analyse par PP de l'espace MT des caractéristiques On effectue une analyse par PP de l'espace MT formé à partir des 20 premières composantes principales représentant 90 pourcent de l'énergie du signal. Les 3 premières projections bidimensionnelles obtenues sont présentées dans la figure 6.7 avec les distributions non-gaussiennes associées.

6.1.3 Modélisation de l'espace des caractéristiques

Notre intérêt est maintenant porté sur la discrimination des structures de caractéristiques des espaces TL et de l'espace MT de dimension réduite, ainsi que sur l'inférence

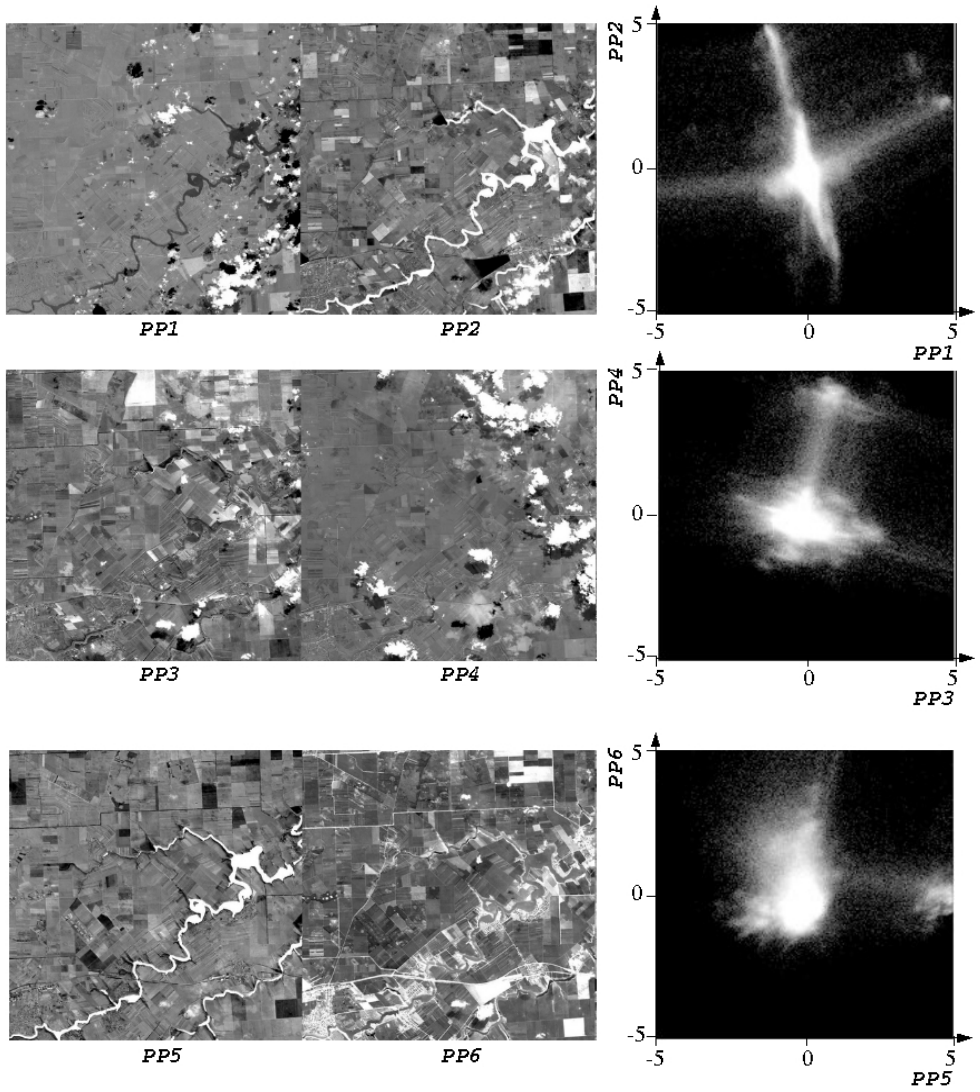


FIG. 6.7 – *3 premières projections bidimensionnelles extraites par PP. Projections représentées dans l'espace image (à gauche) et dans l'espace des caractéristiques (à droite).*

de modèles qui décrivent ces structures. Notre seule connaissance a priori de ces espaces est qu'ils contiennent des caractéristiques possédant une distribution multimodale. En effet, nous ignorons le nombre de structures présentes dans les différents espaces TL et l'espace MT ainsi que la distribution de ces structures. On cherche donc un moyen pour discriminer un nombre inconnu de structures, et pour apprendre la distribution relative à chaque structure. En d'autres termes on souhaite apprendre les paramètres d'un modèle de mélange constitué d'un nombre indéterminé de distributions pour chaque espace. On appellera cluster chaque structure associée à une distribution du mélange.

Pour résoudre ce problème, *Wong et Edward (1993)* proposent l'algorithme de la fonte (melting). Cet algorithme utilise un critère de maximum d'entropie (c.f. § 4.3.1) et une fonction de coût quadratique pour produire une hiérarchie de modèles de mélange liée par un paramètre variable assimilable à une température. On obtient ainsi une modélisation hiérarchique des structures qui sert à l'inférence du nombre de composantes et des paramètres du modèle de mélange. Mais la procédure de la fonte possède une complexité calculatoire prohibitive pour un nombre trop important de points caractéristiques.
Des efforts ont été effectués pour réduire cette complexité. On propose dans *(Ciucu et al., 2003)* une adaptation de la technique de fusion des gradients développée par *Warwick et al. (1996)* pour l'estimation rapide du nombre de clusters et des paramètres des clusters de distributions supposées gaussiennes. On note que seule, cette procédure déterministe d'estimation rapide des clusters est insuffisante car elle présente une sensibilité trop importante à la quantification non-uniforme des caractéristiques dans les espaces TL. Cette procédure servira donc à l'initialisation de l'algorithme de la fonte, accélérant ainsi sa convergence. Mais le coût calculatoire d'une telle modélisation demeure élevée. Toutefois, cette approche nous permet une exploration hiérarchique des structures de caractéristiques à différentes échelles spatiale en relation avec leurs représentations dans l'espace image.
Une solution intéressante pour l'estimation d'un modèle de mélange a été développée par *Lee et al. (2000b)*. La méthode se base sur l'estimation d'un mélange de distributions non-gaussiennes par une ACI.

Pour parvenir à la modélisation souhaitée, on introduit le modèle de mélange de gaussiennes pour lequel on infère le nombre et les paramètres des gaussiennes par le principe MDL.

6.1.3.1 Inférence MDL d'un modèle de mélange de gaussiennes

Afin d'apprendre la distribution mutimodale des caractéristiques dans les espaces TL et l'espace MT de dimension réduite, on utilise un modèle de mélange de gaussiennes $\mathcal{M}_{GM}$. Comme nous l'avons déjà précisé ce modèle est bien adapté à l'estimation de distribution multimodale et permet une approximation efficace d'une distribution pour laquelle aucun a priori n'est disponible. Ce modèle est donc approprié pour l'approximation des distributions multimodales de caractéristiques de différentes natures (spectrales, de texture, etc).
Ce modèle a précédemment été défini dans la section 4.4.3 et l'on conserve les notations employées dans celle-ci. La variable aléatoire Ψ est modélisée par un mélange de gaus-

siennes constitué par K gaussiennes de dimension l. Les paramètres du mélange de gaussiennes servent à discriminer des clusters $\boldsymbol{C}$. En conséquence, dans cette section on fera un amalgame dans les notations entre les clusters et les paramètres $\{\pi_k, M_k, A_k\}$ de poids, de moyennes et de covariance du mélange de gaussiennes . La vraisemblance du modèle est donnée par

$$p(\psi \mid \boldsymbol{C}) = \sum_{k=1}^{K} \pi_k p(\psi \mid M_k, A_k) \ , \ \sum_{k=1}^{K} \pi_k = 1, \tag{6.14}$$

où

$$p(\psi \mid M_k, A_k) = \frac{|A_k^{-1}|^{1/2}}{(2\pi)^{l/2}} \exp\{-\frac{1}{2}(\psi - M_k)^T A_k^{-1}(\psi - M_k)\}. \tag{6.15}$$

Cette famille de modèles paramétriques possède une complexité variable dépendant du nombre de gaussiennes K. L'inférence d'un modèle de mélange de gaussiennes a suscité et suscite encore beaucoup d'intérêt. En employant des approximations de l'évidence bayésienne, *Cheeseman (1988)* dériva un algorithme, appelé "Autoclass", capable d'inférer complètement un mélange de gaussiennes. Plus tard, avec l'apparition du principe MDL et MML employant des longueurs de description en 2-parties (c.f. § 4.3.3.2) et des démonstrations sur l'équivalence des méthodes (c.f. § 4.3.4), des modélisations similaires à l'approche bayésienne ont été développées. Ces méthodes présentent l'avantage d'avoir une complexité calculatoire moindre. *Bischof et al. (1999)* étudièrent dans un contexte MDL l'inférence d'un clustering où chaque cluster est une gaussienne indépendante des autres, possédant des composantes décorrélées et des variances prédéterminées. *Wallace et Dowe (1999)* développèrent l'algorithme "Snob", qui fait l'inférence MML d'un mélange de gaussiennes toujours décorrélées mais aux variances variables.
Dans la lignée de ces modélisations et pour parvenir à une modélisation équivalente à l'algorithme Autoclass mais possédant une complexité calculatoire réduite, on emploie le principe MDL avec une longueur de description en 2-parties qui approxime l'évidence bayésienne d'un modèle de mélange de gaussiennes. C'est à dire qu'on fait l'inférence du nombre de composantes du mélange, des matrices pleines de covariances, des vecteurs de moyenne et des pondérations associées.

Longueur de description en 2-parties pour les modèles de mélange. La longueur de description en 2-parties d'un modèle de mélange est analogue à la longueur de description d'une famille de modèles paramétriques, qui a été détaillée dans la section 4.3.3.2. La seule différence se situe dans l'encodage en plusieurs étapes du modèle. Décrivons donc la longueur de description en 2-parties d'un modèle de mélange. L'approche est la suivante.

Partie 1. La première partie de la longueur de description $-CodeUniv(\boldsymbol{C})$ représente la longueur de code nécessaire à l'encodage du modèle. Celle-ci comprend les longueurs pour décrire :
L_0 , le nombre de distributions du mélange ou clusters ;
L_1 , le nombre de points associé à chaque distribution ;
L_2 , les paramètres relatifs à chaque distribution du modèle ;

L_3 , l'appartenance des points aux clusters.

Partie 2. La deuxième partie de la longueur de description $-CodeUniv(\boldsymbol{\psi^n} \mid \boldsymbol{C})$ est la longueur de code
L_4 , nécessaire à l'encodage des données connaissant le modèle.
Cette quantité est approximée par la quantité $-log\, p(\boldsymbol{\psi^n} \mid \boldsymbol{C})$, et est représentative de la vraisemblance, des paramètres C du modèle de mélange en fonction des données $\boldsymbol{\psi^n} = \{\psi^1, ..., \psi^n\}$, connaissant la complexité K du modèle de mélange $\mathcal{M}_{GM}$.

Soulignons que le dernier terme de la première partie, décrivant l'appartenance des points à leurs clusters, devient rapidement la quantité dominante pour un nombre de points suffisamment grand. Ce terme contrebalance la deuxième partie de la longueur de description qui représente la vraisemblance.
Remarquons par ailleurs que le terme d'appartenance L_3 pourra être soit probabiliste soit déterministe. Dans le cas probabiliste, on autorise les recouvrements partiels des clusters, ce qui permet la modélisation d'un mélange quelconque. Dans le cas déterministe, on se situe dans le cas particulier de clusters qui n'empiètent pas l'un sur l'autre, et on se rapproche ainsi du cas de la quantification de vecteurs.

Application au modèle de mélange de gaussiennes On applique maintenant cette décomposition pour calculer la longueur de description en 2-parties d'un mélange de gaussiennes. On se place dans le formalisme MDL qui emploie des a priori non-informatifs, tels que la distribution uniforme.

- Pour coder le nombre entier $K \in \mathbf{N}$ de gaussiennes du mélange, avec une distribution uniforme pour K pris dans un intervalle d'entier $K \in [1, K_{max}]$, d'après le théorème du codage de Shannon (c.f. § 4.2.3), on a

$$L_0 = log_2(K_{max}). \tag{6.16}$$

- La longueur L_1, représente la longueur de code nécessaire à l'encodage des nombres $\{n_k, k \in [1, K]\}$ de points qui sont associés aux gaussiennes $\{G_k, k \in [1, K]\}$. Cette longueur, calculée en employant un a priori uniforme sur l'intervalle d'entier $n_k \in [1, n]$ où n désigne le nombre total de points, est selon Shannon donnée par

$$L_1 = K log_2(n). \tag{6.17}$$

- Pour chaque gaussienne G_k, le nombre de bits nécessaire à l'encodage de son vecteur moyenne M_k estimé et de sa matrice de covariance A_k estimée est calculé en utilisant une distribution de probabilités uniforme. On est donc ici dans le cas continu et, en conséquence, pour coder chaque paramètres du vecteur et de la matrice de covariances de la gaussienne G_k l dimensionnelle, on a besoin d'une précision $\delta_k = 1/\sqrt{n_k}$ (c.f. § 4.3.3.1). On doit donc calculer les n_k. L'appartenance probabiliste d'un point ψ_i à la gaussienne G_k est définie par

$$\frac{\pi_k . p(\psi_i \mid M_k, A_k)}{p(\psi_i \mid C)}.$$

Mais lorsque l'on calcule l'espérance du nombre n_k de points associés à la gaussienne G_k, on trouve simplement $n_k = \pi_k * n$. On trouve donc que la longueur de code L_2 pour coder les paramètres de l'ensemble des gaussiennes de dimension l, est d'après le théorème du codage de Shannon

$$L_2 = \sum_{k=1}^{K} \frac{l(l+1)}{2} log_2(n\pi_k). \tag{6.18}$$

- Pour décrire l'appartenance des points à leurs gaussiennes associées, la longueur de code nécessaire donnée par le théorème du codage de Shannon est

$$L_3 = \sum_{k=1}^{K} -n\pi_k log_2 \pi_k. \tag{6.19}$$

- Enfin, dans le cas de distributions gaussiennes G_k quelconques, la longueur de code nécessaire pour coder les points $\boldsymbol{\psi^n} = \{\psi^1, ..., \psi^n\}$ en connaissance de la complexité et des paramètres du modèle de mélange, est approximée par l'opposé du logarithme de la vraisemblance. Cette longueur de code est donnée par

$$L_4 = \sum_{i=1}^{n} \sum_{k=1}^{K} \Big(-\frac{1}{2} log_2 \frac{\pi_k^2 |A_k^{-1}|}{2\pi^l} + \frac{1}{ln(2)} (\psi^i - M_k)^t A_k^{-1} (\psi^i - M_k) \Big). \tag{6.20}$$

En conséquence en additionnant les différentes contributions, on obtient la longueur de description en 2-parties pour un modèle de mélange de gaussiennes, qui permet l'encodage de données $\boldsymbol{\psi^n}$. Elle est donc définie par l'expression

$$\begin{aligned} L_{2P}(\boldsymbol{\psi^n}) &= log_2(K_{max}) + K log_2(n) + \sum_{k=1}^{K} \frac{l(l+1)}{2} log_2(n\pi_k) \\ &- \sum_{k=1}^{K} \Big[n\pi_k log_2 \pi_k - \sum_{i=1}^{n} \frac{1}{2} \Big(-log_2 \frac{\pi_k^2 |A_k^{-1}|}{2\pi^l} + \frac{1}{ln(2)} (\psi^i - M_k)^t A_k^{-1} (\psi^i - M_k) \Big) \Big] \end{aligned} \tag{6.21}$$

On remarque que l'appartenance des points aux différentes gaussiennes du mélange s'exprime ici en terme de probabilités par la fonction pondératrice formée par les π_k. Cette appartenance partielle des points aux gaussiennes permet le recouvrement des gaussiennes dans l'espace. Dans le cas particulier où l'appartenance est totale et les recouvrements des gaussiennes du mélange sont négligés, la longueur de description en 2-parties possède des longueurs L_0,L_1,L_2 et L_3 inchangées mais une longueur L_4 égale à

$$\sum_{k=1}^{K} \sum_{\psi_i \in G_k} \Big(-\frac{1}{2} log_2 \frac{|A_k^{-1}|}{2\pi^l} + \frac{1}{ln(2)} (\psi^i - M_k)^t A_k^{-1} (\psi^i - M_k) \Big). \tag{6.22}$$

6.1.3.2 Minimisation de la longueur de description

Le but de l'estimation est la minimisation de la longueur de description en 2-parties $L_{2P}(\boldsymbol{\psi^n})$, qui par le principe MDL permet l'estimation du nombre de gaussiennes $\widehat{K}$,

de leurs paramètres de moyenne $\widehat{M}_k$, de covariance $\widehat{A}_k$, et de poids $\widehat{\pi}_k$. L'énumération de toutes les configurations pour effectuer cette minimisation n'est pas faisable. On a donc recours à une procédure d'optimisation qui après une phase d'initialisation (phase d'initialisation), emploie une optimisation par une méthode de plus forte descente et une optimisation par l'algorithme Espérance-Maximisation (EM) (c.f. § 4.5). L'optimisation par l'algorithme EM constitue la phase d'adaptation du modèle aux données qui permet l'estimation des matrices de covariance, des vecteurs de moyenne et de pondération pour un modèle de complexité déterminé (phase d'adaptation) tandis que l'optimisation par la méthode de plus forte descente constitue la phase de sélection de la complexité du modèle (phase de sélection). Ces deux phases d'optimisation sont itérées jusqu'à convergence de l'algorithme vers un des minima de $L_{2P}(\boldsymbol{\psi^n})$ (phase de convergence). En effet, on a vu que la convergence des algorithmes EM et de plus forte descente est assurée, mais que la convergence vers un maximum global de la fonction de vraisemblance n'est pas garantie. Avant de décrire l'algorithme d'optimisation globale, détaillons dans un premier temps la procédure itérative de maximisation de la vraisemblance par l'algorithme EM, et dans un deuxième temps, la variation de la longueur $L_{2P}(\boldsymbol{\psi^n})$ employée par la méthode de plus forte descente. L'itération courante de l'algorithme EM sera notée (p) tandis que l'itération courante de l'algorithme de plus forte descente sera notée (q).

Estimation de la vraisemblance du modèle par l'algorithme EM. A l'itération (q), pour adapter un modèle de mélange de $K^{(q)}$ gaussiennes aux données $\boldsymbol{\psi^n} = \{\psi^1, ..., \psi^n\}$ et ainsi maximiser la vraisemblance $p(\boldsymbol{\psi^n} \mid C, \mathcal{M}_{GM})$, on emploie l'algorithme EM décrit dans la section 4.5. L'algorithme itératif est constitué à l'itération $^{(p)}$ par

– *l'étape "E" :*

$$h_k^{(p)}(\psi_i) = \frac{\pi_k^{(p)} p(\psi_i \mid M_k^{(p)}, A_k^{(p)})}{\sum_{j=1}^{K^{(q)}} \pi_j^{(p)} p(\psi_i \mid M_j^{(p)}, A_j^{(p)})} \tag{6.23}$$

– *l'étape "M" :*

$$\pi_k^{(p+1)} = \frac{\sum_{i=1}^{n} h_k^{(p)}(\psi_i)}{n} \tag{6.24}$$

$$M_k^{(p+1)} = \frac{\sum_{i=1}^{n} h_k^{(p)}(\psi_i)\psi_i}{\sum_{i=1}^{n} h_k^{(p)}(\psi_i)} \tag{6.25}$$

$$A_k^{(p+1)} = \frac{\sum_{i=1}^{n} h_k^{(p)}(\psi_i)(\psi_i - M_k^{(p+1)})(\psi_i - M_k^{(p+1)})^T}{\sum_{i=1}^{n} h_k^{(p)}(\psi_i)} \tag{6.26}$$

Dans le cas particulier où le recouvrement des gaussiennes est négligé (c.f. section précédente), la probabilité d'appartenance d'un point ψ_i à la gaussienne k est égale à 1 si la gaussienne en question maximise la probabilité du point et égale à 0 dans le cas contraire. Dans ce cas, on estime donc dans l'étape "M" les paramètres des gaussiennes sans considérer la possible appartenance des points aux gaussiennes voisines. En conséquence, l'algorithme EM possède à l'itération (p) la même étape "M" mais l'étape "E" est différente. Dans ce

cas l'étape E devient

$$h_k^{(p)}(\psi_i) = \begin{cases} 1 & \text{si } p(\psi_i \mid M_k^{(p)}, A_k^{(p)}) = \max_{j \in K^{(q)}} p(\psi_i \mid M_j^{(p)}, A_j^{(p)}) \\ 0 & \text{dans le cas contraire} \end{cases} \tag{6.27}$$

Pour un certain nombre de gaussiennes $K^{(q)}$, en itérant cet algorithme jusqu'à convergence, on estime par maximum de vraisemblance les paramètres du modèle de mélange $\widehat{C}_{ML}^{(q)}$.

Variation de longueur de description. Une optimisation par plus forte descente utilise une mesure qui évalue la plus forte pente de la fonction à minimiser. Dans notre cas, à l'itération (q), nous utilisons la variation de la longueur de description lorsqu'une gaussienne $G_{k'}$ est supprimée du modèle de mélange dont les paramètres sont $\widehat{C}_{ML}^{(q)}$. Cette variation $\Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n}))$ se décompose en 4 termes

$$\Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n})) = \Delta_{G_{k'}}^{(q)}(L_1) + \Delta_{G_{k'}}^{(q)}(L_2) + \Delta_{G_{k'}}^{(q)}(L_3) + \Delta_{G_{k'}}^{(q)}(L_4) \tag{6.28}$$

correspondant aux variations des longueurs de L_1, L_2, L_3 et L_4 après la suppression de la gaussienne $G_{k'}$ à l'itération $^{(q)}$. On note que la variation de la longueur L_0 est nulle.
$\Delta_{G_{k'}}^{(q)}(L_1)$ est égale à la réduction de la longueur de code obtenue par la suppression du codage du nombre $n_k^{'}$. Cette quantité est donc définie par

$$\Delta_{G_{k'}}^{(q)}(L_1) = -log_2(n). \tag{6.29}$$

$\Delta_{G_{k'}}^{(q)}(L_2)$ est égale à la réduction de la longueur de code obtenue par la suppression du codage des paramètres de la gaussiennes $G_{k'}$. Cette quantité est donc définie par

$$\Delta_{G_{k'}}^{(q)}(L_2) = -\frac{l(l+1)}{2} log_2(n\pi_{k'}). \tag{6.30}$$

$\Delta_{G_{k'}}^{(q)}(L_3)$ est égale à la réduction de la longueur de code obtenue par la suppression du codage des indices d'appartenances des points associés à la gaussienne $G_{k'}$. Cette quantité est donc définie par

$$\begin{aligned} \Delta_{G_{k'}}^{(q)}(L_3) &= -\sum_{j=1, j\neq k'}^{K^{(q)}} n(\pi_j + \pi_{jk'}) log_2(\pi_j + \pi_{jk'}) + \sum_{j=1}^{K^{(q)}} n\pi_j log_2 \pi_j \\ &= n\pi_{k'} log_2 \pi_{k'} - \sum_{j=1, j\neq k'}^{K^{(q)}} n\big((\pi_j + \pi_{jk'}) log_2(\pi_j + \pi_{jk'}) - \pi_j log_2 \pi_j\big). \end{aligned} \tag{6.31}$$

où $\pi_{jk'}$ est le poids de G_j, après suppression de $G_{k'}$, estimé par maximum de vraisemblance à l'itération $(q+1)$.
Enfin, $\Delta_{G_{k'}}^{(q)}(L_4)$ est égale à l'augmentation de la longueur de code causée par la diminution de la vraisemblance du modèle. Cette quantité est donc définie par

$$\Delta_{G_{k'}}^{(q)}(L_4) = \sum_{i=1}^{n} log_2 \frac{p(\psi_i \mid C_{ML}^{(q)})}{p(\psi_i \mid C_{ML}^{(q+1)})} \tag{6.32}$$

où $C_{ML}^{(q+1)}$ sont les paramètres du modèle, après suppression de $G_{k'}$, estimés par maximum de vraisemblance à l'itération $(q+1)$. On doit donc en théorie faire converger l'algorithme EM pour l'estimation de $C_{ML}^{(q+1)}$. Toutefois, en pratique, parce que l'on doit seulement comparer les variations engendrées par la suppression des différentes gaussiennes du mélange, quelques itérations de l'algorithme suffisent et l'on ne fait pas converger l'algorithme. Par contre, lorsque après chaque simulation de suppression de gaussienne, on observe seulement des augmentations de la longueur de code, alors pour s'assurer de la convergence de l'algorithme, plusieurs itérations de l'algorithme EM seront nécessaires pour estimer $C_{ML}^{(q+1)}$.
En additionnant les termes, on trouve donc

$$\Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n})) = -log_2(n) - \frac{l(l+1)}{2} log_2(n\pi_{k'}) + n\pi_{k'} log_2 \pi_{k'} \quad (6.33)$$
$$- \sum_{j=1, j\neq k'}^{K^{(q)}} n\big((\pi_j + \pi_{jk'}) log_2(\pi_j + \pi_{jk'}) - \pi_j log_2 \pi_j\big) + \sum_{i=1}^{n} log_2 \frac{p(\psi_i \mid C_{ML}^{(q)})}{p(\psi_i \mid C_{ML}^{(q+1)})}$$

Dans le cas on l'on néglige les recouvremente des gaussiennes dans l'espace des caractéristiques, les termes $\Delta_{G_{k'}}^{(q)}(L_1)$, $\Delta_{G_{k'}}^{(q)}(L_2)$ et $\Delta_{G_{k'}}^{(q)}(L_3)$ restent inchangés. Par contre le terme $\Delta_{G_{k'}}^{(q)}(L_4)$ est différent. Pour alléger les équations, introduisons des notations. On note par $p_j(\psi_i)$ la distribution gaussienne $p(\psi \mid M_j, A_j)$ du modèle de mélange $p(\psi \mid C_{ML}^{(q)})$. On note par $p_{jk'}(\psi_i)$ la distribution gaussienne $p(\psi \mid M_j, A_j)$ du modèle de mélange $p(\psi \mid C_{ML}^{(q+1)})$ qui a été estimé après suppression de $G_{k'}$. Enfin, on note par $p_{k'k'}(\psi_i)$ la distribution de la gaussienne la plus probable qui est associée à ψ_i après suppression de $G_{k'}$. Muni des ces notations, dans ce cas particulier, la variation de la deuxième partie de la longueur de description $\Delta_{G_{k'}}^{(q)}(L_4)$ devient

$$\sum_{\psi_i \in G_{k'}} log_2 \frac{p_{k'}(\psi_i)}{p_{k'k'}(\psi_i)} + \sum_{j=1, j\neq k'}^{K^{(q)}} \sum_{\psi_i \in G_j} log_2 \frac{p_j(\psi_i)}{p_{jk'}(\psi_i)}. \quad (6.34)$$

La première somme est la variation de la longueur de code relative à l'encodage des points appartenant à la gaussienne supprimée tandis que le deuxième terme est la variation relative aux autres points.

Estimation de la complexité et des paramètres du modèle. Pour estimer conjointement la complexité du modèle, c'est à dire le nombre de gaussiennes K, et les paramètres du modèle par maximum de vraisemblance, on minimise la longueur de description $L_{2P}(\boldsymbol{\psi^n})$ en alliant une méthode de plus forte descente avec l'algorithme EM. L'idée est de partir d'un nombre initial $K^{(0)}$ élevé de gaussiennes et de réduire progressivement ce nombre, en maximisant la réduction de la longueur de description, à chaque suppression. Lorsque l'on atteindra le stade où l'on ne pourra plus supprimer de gaussiennes du mélange sans augmenter la longueur de description, on aura atteint un des minima de $L_{2P}(\boldsymbol{\psi^n})$ et l'on aura estimé la complexité et les paramètres du modèle de mélange.

Plus précisément, la procédure d'optimisation est la suivante.

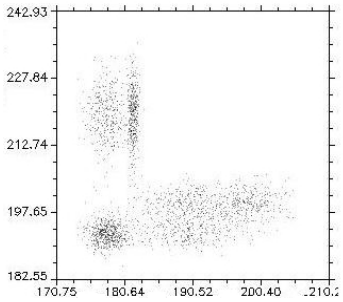
1-Initialisation : A l'itération (0), Un modèle de mélange avec un nombre $K^{(0)}$ de gaussiennes et des paramètres $C^{(0)}$ est généré. L'initialisation est faite par la répartition aléatoire des $K^{(0)}$ composantes du mélange dans l'espace des caractéristiques en fonction pour chaque dimension d'une distribution gaussienne possédant une moyenne et une variance respectivement égale à la moyenne et la variance statistique des données.

2-Adaptation : A l'itération (q), on estime d'abord par l'algorithme EM le maximum de vraisemblance des paramètres $\widehat{C}_{ML}^{(q)}$ du modèle de mélange pour le nombre courant $K^{(q)}$ de gaussiennes.

3-Sélection : Toujours à l'itération (q), on supprime la gaussienne $G_{k'}$ entraînant la plus importante réduction de la longueur de description $\Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n}))$ puis l'on retourne à la phase 2 de l'algorithme pour l'itération $(q+1)$. Si aucune réduction de la longueur de description est observée, c'est à dire si $\forall G_{k'},\ \Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n})) < 0$, alors on ne supprime aucune gaussienne et l'on passe à la phase 4.

4-Convergence : Lorsque à une itération (q_fin), plus aucune réduction de la longueur de description n'est observée, l'algorithme itératif est arrêté. On obtient alors le nombre de gaussiennes mélangées $\widehat{K} = K^{(q\_fin)}$ avec les paramètres du modèle de mélange $\{\widehat{M}_k, \widehat{A}_k, \widehat{\pi}_k\}$ donnés par $\widehat{C}_{ML}^{(q\_fin)}$.

Les résultats présentés dans cette thèse sont issus d'une modélisation par mélange de gaussiennes où les recouvrements partiels des gaussiennes sont négligés. Contrairement au cadre généraliste, ce cas particulier est restrictif et ne permet pas toujours l'approximation d'une distribution quelconque. Afin d'utiliser des hypothèses valides quelle que soit la distribution à modéliser, on se doit donc d'implémenter le modèle général. Une telle implémentation, bien que facilement réalisable, n'a pas aboutie faute de temps. Par ailleurs, la variation de longueur de code relative à l'encodage des nombres de points associés aux différents éléments du mélange (terme $\Delta_{G_{k'}}^{(q)}(L_1)$) a été négligée et la variation relative à l'encodage des vecteurs de moyenne et des matrices de covariance (terme $\Delta_{G_{k'}}^{(q)}(L_2)$) a été approximée par le terme $-l(l+1)$. Enfin, lorsque l'on simule la suppression d'une gaussienne, on se limite à une seule itération de l'algorithme EM. Comme déjà remarqué, cette approche est suffisante lorsque des réductions de $\Delta_{G_{k'}}^{(q)}(L_{2P}(\boldsymbol{\psi^n}))$ sont observées, mais peut être insuffisante pour s'assurer de la convergence de l'algorithme. Cette insuffisante peut aboutir à une surestimation de la complexité du mélange de gaussiennes.
Cependant, ce cas particulier de modélisation par mélange de gaussiennes est capable de représenter une distribution multimodale dans la plupart des cas. Pour évaluer l'efficacité de la modélisation, des expérimentations ont été effectuées avec des données générées par un mélange de gaussiennes bidimensionnelles. La figure 6.8 présente les jeux de données de taille n variable dans un espace de dimension accrue par l'ajout de composantes de bruit gaussien ou/et de composantes corrélées. Les résultats présentés dans cette même figure démontrent que, pour un nombre de réalisations suffisamment grand, cette modélisation est capable d'estimer correctement le véritable nombre de gaussiennes et les paramètres du



Paramètres du modèle					
Cluster	M1	M2	σ1	σ2	πi
A	200	200	3	3	12.5
B	190	200	5	3	20.8
C	190	193	5	2.5	16.7
D	178	193	2	2.5	16.7
E	178	220	2	7	20.8
F	182	220	0.5	7	12.5

n	Nombre de composantes bruitées	Nombre de composantes corrélées	Nombre de clusters retrouvés	Clusters retrouvés					
				A	B	C	D	E	F
100x100	0	0	6	x	x	x	x	x	x
100x100	0	1	5	x			x	x	x
100x100	0	2	4					x	x
125x125	0	2	6	x	x	x	x	x	x
100x100	1	0	3				x		
125x125	1	0	6	x	x	x	x	x	x
125x125	2	0	3						x
500x250	4	0	6	x	x	x	x	x	x

FIG. 6.8 – *Evaluation de l'algorithme d'inférence par MDL d'un modèle de mélange de gaussiennes. En haut : l'histogramme bidimensionnel de 10000 réalisations d'un modèle de mélange (à gauche) et les paramètres de ce modèle (à droite). En bas : évaluation de l'influence de la dimension de l'espace et du nombre de réalisations sur l'inférence du modèle de mélange.*

modèle de mélange. Basé sur les résultats de cette étude, on considère que cet algorithme est fiable et permet une modélisation efficace.

6.1.4 Clustering et classifications des caractéristiques

On procède donc à l'inférence d'un modèle de mélange de gaussiennes pour les différentes représentations du signal. D'une part, par la modélisation des caractéristiques de l'espace MT réduit $\mathcal{E}_{\mathcal{P}(\theta_1,...,\theta_d)}$, on obtient une collection de clusters MT. Afin de représenter ces clusters dans l'espace MT $\mathcal{E}_{\theta_{t_1},...,\theta_{t_d}}$, on applique aux clusters la projection inverse $\mathcal{P}^{-1}$. On forme donc une collection de clusters MT C^j_{MT} avec $j \in [1, K_{MT}]$ et où K_{MT} est le nombre estimé de clusters. Le processus de modélisation produit également une partition dans l'espace euclidien $\mathcal{E}_r$. Les classes engendrées par ce processus sont notées $C^j_{MT}(r)$.
D'autre part, la modélisation des caractéristiques de chaque espace TL, mène à la création de d collections de clusters dans les espaces TL $\{\mathcal{E}_{\theta_{t_i}};\ i = 1, ..., d\}$. Ces clusters sont qualifiés de TL et notés $\{C^j_{t_i}; t_i \in [1, d]\}$ avec $j \in [1, K_{t_i}]$ et où K_{t_i} représente le nombre de clusters estimés au temps t_i. Dans l'espace $\mathcal{E}_r$, les classes correspondantes sont notées $C^j_{t_i}(r)$.

En modélisant pour chacune des 38 images de la STIS, les caractéristiques TL prises dans une fenêtre spatiale de 800x800 pixels, on obtient une série de collection de clusters.

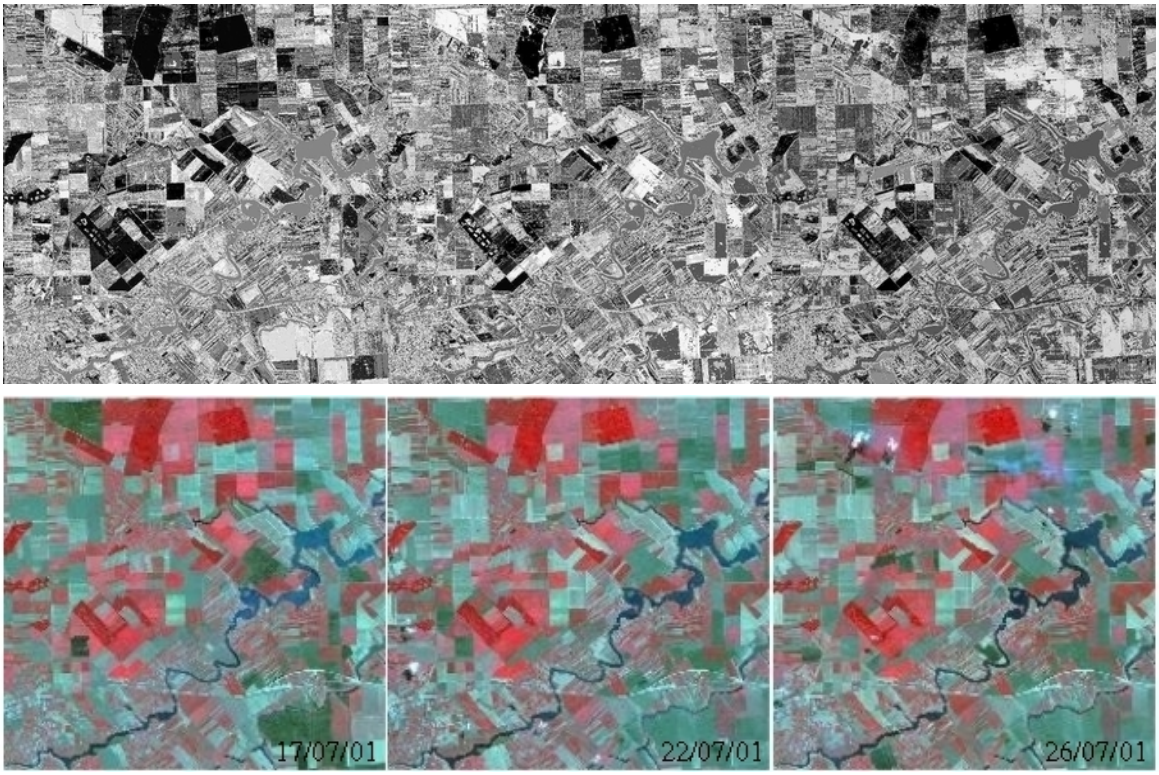


FIG. 6.9 – *Images du 17, 22 et 26 mai 2001 (en bas) et leurs classifications associées (en haut) contenant respectivement 82, 81 et 68 classes.*

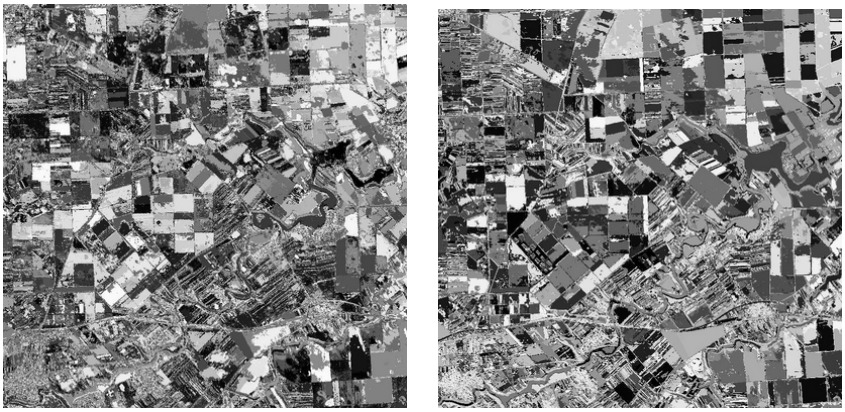


FIG. 6.10 – *Classification MT postérieure à une analyse par PP (droite) et postérieure à une analyse par ACP (gauche), contenant respectivement 160 et 138 classes.*

Pour ces caractéristiques TL, l'initialisation du modèle de mélange est effectuée avec la collection de clusters trouvée au temps précédent. De plus, cette collection est complétée par des clusters supplémentaires pour former une collection de cardinalité supérieure à 120. Ce chiffre est cohérent avec les études menées par *Datcu et Seidel (2002)* sur la complexité des caractéristiques d'images satellitaires. La figure 6.9 présente les classifications de 3

images consécutives de la STIS.
La modélisation MT est effectuée sur les caractéristiques de dimension réduite provenant de la même zone spatiale. On constitue l'espace MT à partir des 3 projections bidimensionnelles extraites par PP ne présentant pas de nuages. Plus de 200 clusters servent à l'initialisation de la procédure. L'algorithme infère un modèle de mélange constitué de 160 clusters. A titre comparatif, une modélisation est parallèlement effectuée avec un espace MT formé à partir des 6 premières composantes fournies par une ACP. On obtient un nombre de gaussiennes inférieur à la modélisation précédente. Ce résultat atteste de l'efficacité de l'algorihme de PP à révéler les distributions multimodales. La figure 6.10 présente les deux classifications MT obtenues.

6.2 Inférence de trajectoires de structures spatio-temporelles

L'objectif est maintenant d'obtenir une modélisation du signal $x_r(t, \theta)$ dans l'espace des caractéristiques évolutives $\mathcal{E}_{t,\theta}$ par l'inférence de trajectoires de clusters dynamiques.
Les clusters MT $\{C^j_{MT}; j = 1, ..., K_{MT}\}$ du niveau hiérarchique précédent peuvent être perçus comme des regroupements de caractéristiques possédant des évolutions temporelles similaires. Les clusters MT sont donc assimilables à des clusters dynamiques. Pour obtenir des trajectoires de clusters dynamiques à partir des clusters MT, on essaye donc de représenter les clusters de l'espace MT des caractéristiques $\mathcal{E}_{\theta_1,...,\theta_d}$ dans l'espace des caractéristiques évolutives $\mathcal{E}_{t,\theta}$.
Pour l'obtention de ces trajectoires, on cherche par ailleurs à introduire l'information complémentaire fournie par les clusters TL $\{C^j_{t_i}; j = 1, ..., K_{t_i}\}$ pour $t_i \in [1, d]$. Or, les espaces TL des caractéristiques $\{\mathcal{E}_{\theta_{t_i}}; i = 1, ..., d\}$ constituent des sous-espaces de l'espace MT des caractéristiques et les différents états de l'espace des caractéristiques évolutives. L'idée consiste donc à formaliser des trajectoires de clusters MT, assimilables à des clusters dynamiques, au travers des clusters TL. De plus, on introduit dans ce processus d'inférence de trajectoires une contrainte spatiale par la prise en compte des recouvrements entre classes TL et classes MT. Comme on le verra par la suite, la formalisation de trajectoires de clusters MT au travers de clusters TL permettra parallèlement l'inférence de classes dynamiques associées aux différentes trajectoires.
On présente dans un premier temps le modèle et le processus d'inférence de trajectoires, puis dans un deuxième temps, on détaille les objets engendrés par ce processus, c'est à dire un graphe caractéristique de trajectoires de clusters dynamiques et des classifications dynamiques associées.

6.2.1 Modélisation des trajectoires de clusters dynamiques

On introduit maintenant le modèle pour l'inférence des trajectoires de clusters dynamiques.
Mais avant de détailler ce modèle on s'attache d'abord à rendre les clusters TL et MT

comparables. A un temps t_i, on doit donc rendre commensurable un cluster C^{j1}_{MT} défini dans l'espace MT $\mathcal{E}_{\theta_{t_1},...,\theta_{t_d}}$ et un cluster $C^{j2}_{t_i}$ défini dans l'espace TL $\mathcal{E}_{\theta_{t_i}}$. Pour ce faire, il suffit simplement de supprimer les composantes $\{\theta_k; k \neq i\}$ du cluster MT. Une telle marginalisation du clusters C^{j1}_{MT} sur t_i aboutit à un cluster MT projeté sur θ_{t_i} qui est noté $C^{j1}_{MT_{t_i}}$. Ce cluster et un cluster TL $C^{j2}_{t_i}$ appartiennent maintenant au même espace TL $\mathcal{E}_{\theta_{t_i}}$.

Les trajectoires des clusters MT ont été étudiées et caractérisées dans *(Héas et al., 2002)*. Ces études nous ont amenées à développer le modèle de trajectoires présenté dans la suite.

Pour la modélisation de l'évolution des clusters, on emploie un modèle de trajectoire de clusters MT noté $\mathcal{M}_T$. Ce modèle est une distribution de probabilités sur une collection $\boldsymbol{C}$ de clusters MT qui est conditionnée par un certain graphe $\mathcal{G}$ de trajectoires, constituées par un certain nombre de clusters TL.
Ce modèle évalue donc la qualité ou la vraisemblance de l'association des clusters MT avec les clusters TL. Pour définir cette vraisemblance $p(\boldsymbol{C} \mid \mathcal{G}, \mathcal{M}_T)$, on fait l'hypothèse (1) que l'on obtient indépendamment les différentes trajectoires des clusters MT. Pour simplifier le modèle, on fait aussi l'hypothèse (2) que conditionnellement à un certain cluster MT, l'association avec un cluster TL est indépendante des associations avec les autres clusters TL (localisés au même temps ou à des temps différents). En notant par $\mathcal{G}_k$ le sous-graphe de $\mathcal{G}$ associé à la trajectoire du cluster MT $(C^k_{MT_{t_i}})$, et en supposant la première hypothèse d'indépendance, on définit le modèle $\mathcal{M}_T$ par la distribution

$$p(\boldsymbol{C} \mid \mathcal{G}, \mathcal{M}_T) = p(C^1_{MT}, ..., C^{K_{MT}}_{MT} \mid \mathcal{G}) \tag{6.35}$$

$$= \prod_{k=1}^{K_{MT}} p(C^k_{MT} \mid \mathcal{G}_k) \tag{6.36}$$

où $\mathcal{G}_k$ est constitué par une collection de clusters TL $\{C^j_{t_i}\}$. En employant la règle de Bayes et la deuxième hypothèse d'indépendance on a

$$p(C^k_{MT} \mid \mathcal{G}_k) = \frac{p(C^k_{MT})p(\{C^j_{t_i}\} \mid C^k_{MT})}{p(\{C^j_{t_i}\})} \tag{6.37}$$

$$= p(C^k_{MT}) \prod_{(i,j),\, C^j_{t_i} \in \mathcal{G}_k} \frac{p(C^j_{t_i} \mid C^k_{MT})}{p(C^j_{t_i})} \tag{6.38}$$

En conséquence, en prenant une distribution uniforme pour $p(C^k_{MT})$ et $p(C^j_{t_i})$, les termes du produit de l'Eq. 6.36 vérifient la relation

$$p(C^k_{MT} \mid \mathcal{G}_k) \propto \prod_{(i,j),\, C^j_{t_i} \in \mathcal{G}_k} p(C^j_{t_i} \mid C^k_{MT}), \tag{6.39}$$

et l'Eq. 6.36 se réécrit sous la forme du double produit

$$p(\boldsymbol{C} \mid \mathcal{G}_k, \mathcal{M}_T) \propto \prod_{k=1}^{K_{MT}} \prod_{(i,j),\, C^j_{t_i} \in \mathcal{G}_k} p(C^j_{t_i} \mid C^k_{MT}). \tag{6.40}$$

On définit les probabilités d'associations en employant la divergence de Kullback-Leibler $D(.,.)$ pour mesurer la similarité entre clusters dans l'espace des caractéristiques, et l'opérateur $card(.)$ qui prend pour argument une classe spatiale et retourne sa population dans le but d'introduire une contrainte de recouvrement spatial. La divergence est une mesure entropique capable de mesurer la similarité de deux distributions et constitue un outil adapté à la sélection de modèle (c.f. 4.3.1). Dans un soucis de simplicité, on choisit d'utiliser un simple rapport pour définir les probabilités d'association

$$p(C_{t_i}^j \mid C_{MT}^k) = \frac{1}{Z} \frac{card(C_{MT}^k(r) \cap C_{t_i}^j(r))}{D(C_{MT_{t_i}}^k, C_{t_i}^j)} \tag{6.41}$$

où Z est une constante de normalisation. Le maximum de la probabilité de vraisemblance (c.f. Eq. 6.40) s'obtient avec des graphes $\mathcal{G}_k$ comprenant tous les clusters TL $C_{t_i}^j$. Mais, comme l'objectif est d'inférer seulement les associations des clusters MT avec les clusters TL les plus probables, on limite la complexité des graphes $\mathcal{G}_k$ par la suppression des clusters TL possédant une probabilité $p(C_{t_i}^j \mid C_{MT}^k)$ inférieure à un certain seuil μ. Par exemple, une valeur de μ égale à 0.9 correspondra à un graphe possédant des probabilités d'associations supérieures à 0.9. En conséquence, les graphes $\mathcal{G}_k$ qui maximisent la vraisemblance $p(\boldsymbol{C} \mid \boldsymbol{\mathcal{G}}_k, \mathcal{M}_T)$ sont simplement ceux qui sont constitués de clusters TL $C_{t_i}^j$ où les probabilités $p(C_{t_i}^j \mid C_{MT}^k)$ sont supérieures à μ. Le seuil μ, déterminant de la complexité des graphes, pourra être ajusté par l'utilisateur dans le niveau suivant de la hiérarchie.
Si l'on veut inférer seulement la trajectoire la plus probable relative à chaque cluster MT C_{MT}^k, alors on conserve à chaque temps t_i, le clusters TL qui maximise sur j la probabilité d'association $p(C_{t_i}^j \mid C_{MT}^k)$.
Notons que le calcul des divergences est analytique car les clusters sont caractérisés par des distributions gaussiennes multivariées. Pour pallier les problèmes de calcul lorsque l'on compare deux gaussiennes de variances très différentes, on approxime les distributions très échancrées par des Diracs. Dans ce cas, on a également recours à un calcul analytique.

En utilisant les clusters MT, on infère donc des graphes de trajectoires de clusters dynamiques qui sont composées de clusters TL. Les graphes possèdent une complexité fonction du paramètre de seuil μ. Ces graphes constituent des modèles pour la représentation du signal $x_r(t, \theta)$ dans l'espace des caractéristiques évolutives.

6.2.2 Graphe de trajectoires et classifications dynamiques

Graphe de trajectoires. La série temporelle d'images $x(r, t, \theta)$ a subi précédemment plusieurs processus de traitement. Par l'application de ces méthodes, on obtient une représentation spécifique $x_r(t, \theta)$ qui sont des graphes modélisant les trajectoires des clusters dynamiques. La chronologie de la série temporelle et l'échantillonnage irrégulier sont des informations stockées dans ces graphes. D'autres informations concernant les trajectoires sont contenues dans les noeuds et les arcs de ces graphes. Un processus de caractérisation des trajectoires engendre une collection d'attributs associés aux noeuds et aux arcs du graphe de trajectoires. La figure 6.11 présente les différents attributs associés

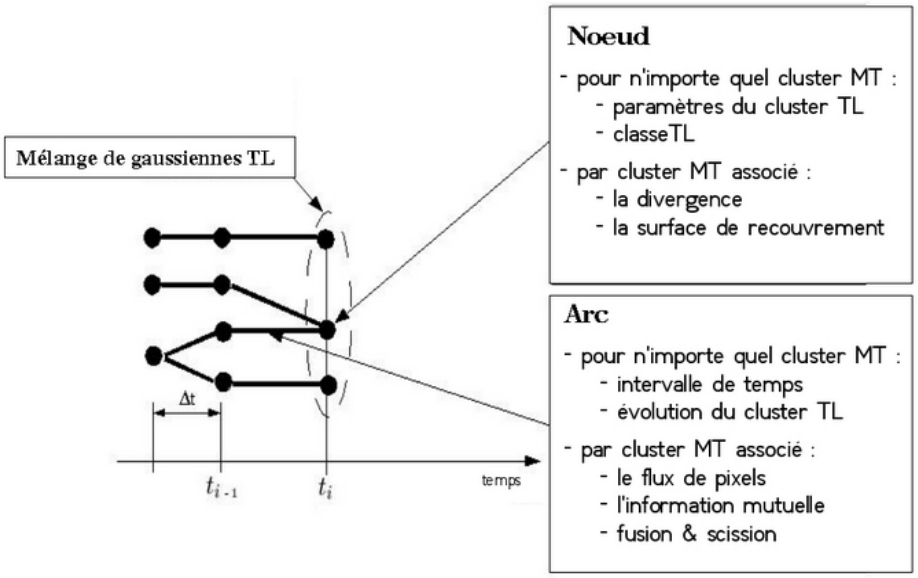


FIG. 6.11 – *Description du graphe de trajectoires de clusters dynamiques.*

au graphe.

Détaillons ces différents attributs et les processus de caractérisation les engendrant.

- Un noeud représente un cluster TL $C^j_{t_i}$ qui est la j-ème composante du mélange de gaussiennes au temps t_i. Ce cluster TL est associé à une collection de clusters MT par les probabilités définies dans l'équation 6.41. Le mélange de gaussiennes au temps t_i est décrit par l'ensemble des clusters TL. Chaque cluster MT C^k_{MT} associé au noeud $C^j_{t_i}$ est caractérisé par une surface de recouvrement de classes $card(C^k_{MT_{t_i}}(r) \cap C^j_{t_i}(r))$, une mesure de divergence et les paramètres de la gaussienne associée au cluster $C^j_{t_i}$. De plus, chaque noeud regroupe un ensemble de points indexés dans le temps et dans l'espace qui peuvent être représentés par la classe $C^j_{t_i}(r)$.

- Un arc du graphe représente l'évolution d'un cluster MT entre deux échantillons temporels. Un arc, reliant 2 clusters TL consécutifs $C^j_{t_i}$ et $C^l_{t_{i-1}}$, et associé à un certain cluster MT $\{C^k_{MT}\}$, est caractérisé par un intervalle de temps Δt_i, un flux de points caractéristiques et des évolutions de paramètres gaussiens TL et MT. Le flux de pixels échangé entre le cluster TL $C^j_{t_i}$ et $C^l_{t_{i-1}}$ est le nombre de pixels en commun dans les deux classes spatiales TL $C^j_{t_i}(r)$ et $C^l_{t_{i-1}}(r)$. De plus, parce qu'on se restreint à une certaine classe MT associée, le flux est déterminé par le nombre de pixels en commun entre les 2 classes précédentes et la classe MT $C^k_{MT}(r)$. La quantification du flux permet l'évaluation des phénomènes de fusion et de scission. Ces phénomènes sont simplement reliés au nombre d'arcs partant et arrivant à un certain noeud. Le changement interne du cluster MT entre deux dates t_{i-1} et t_i peut être quantifié par l'information mutuelle (c.f. § 4.2.2).

Le changement interne à un cluster MT entre deux temps consécutifs est quantifié par l'information mutuelle. En accord avec les notations précédentes, les projections d'un cluster MT C^k_{MT} à des temps consécutifs t_i et t_{i+1} sont notés $C^k_{MT_{t_i}}$ et $C^k_{MT_{t_{i+1}}}$. Ces projections sont des distributions gaussiennes multivariées, qui ne sont autres que les marginales de la distribution d'un cluster MT. La distribution conjointe, notée $C^k_{MT_{t_i \cup t_{i+1}}}$, est obtenue par la projection d'un cluster MT dans l'espace formé à partir des deux composantes de caractéristiques TL aux instants t_i et t_{i+1}. En conséquence, comme les distributions marginales et la distribution conjointe sont gaussiennes, l'information mutuelle peut être calculée analytiquement

$$MI(C^k_{MT_{t_i}}, C^k_{MT_{t_{i+1}}}) = \frac{1}{2} log \frac{|A^k_{t_i}||A^k_{t_{i+1}}|}{|A^k_{t_i \cup t_{i+1}}|} \tag{6.42}$$

où $|A^k_{t_i}|$,$|A^k_{t_{i+1}}|$ et $|A^k_{t_i t_{i+1}}|$ sont respectivement les déterminants des matrices de covariance attachées aux projections gaussiennes des clusters MT $C^k_{MT_{t_i}}$, $C^k_{MT_{t_{i+1}}}$ et $C^k_{MT_{t_i \cup t_{i+1}}}$ *(Kullback, 1966)*.

Classifications dynamiques. Le graphe de trajectoires de clusters dynamiques $x_r(t, \theta)$ est une représentation du signal où la variable spatiale r est cachée. Mais, les indices spatiaux relatifs à chaque point de l'espace de caractéristiques sont accessibles et permettent la génération de représentations dans l'espace image. En effet, on peut associer aux K_{MT} différentes trajectoires des clusters MT, K_{MT} différentes représentations dans l'espace spatio-temporel formé à partir des variables r et t. On appelle ces représentations des classifications dynamiques. Chacune d'elles est composée à chaque temps $t_i \in [1, d]$ par une combinaison particulière de classes TL $C^j_{t_i}(r)$. Les classes TL présentes dans la classification dynamique relative à la k-ème classe MT $C^k_{MT}(r)$, sont celles qui correspondent à des clusters TL dans le graphe de trajectoires $\mathcal{G}_k$ associé.
Ces classifications dynamiques constituent le signal $x_\theta(r, t)$. Celui-ci contient l'information spatio-temporelle qui est absent du signal $x_r(t, \theta)$ représenté dans l'espace des caractéristiques évolutives. Ensemble, ces représentations décrivent objectivement l'évolution des caractéristiques et l'évolution spatiale de la STIS $x(r, t, \theta)$.
Notons que l'on pourrait facilement fusionner ces deux représentations en extrayant des caractéristiques géométriques à partir des classifications dynamiques. Ainsi, par cette extraction d'information spatiale, on pourrait dériver de nouveaux attributs qui compléteraient le graphe. Le graphe caractériserait alors à lui seul le signal $x(r, t, \theta)$.

La figure 6.12 présente des exemples de graphes de trajectoires et de classifications dynamiques associées à deux classes MT. Les liens entre ces objets sont par ailleurs mis en évidence. Une analyse visuelle des classes MT, des graphes de trajectoires, des classifications dynamiques et de la SITS originale, permet la compréhension du phénomène de labours progressifs.

- On observe d'abord la répercussion du labourage d'un champ effectué en deux parties sur le graphe de trajectoires. Le motif engendré se caractérise par une scission puis une fusion de clusters accompagnées de ruptures brusques des trajectoires (classe MT

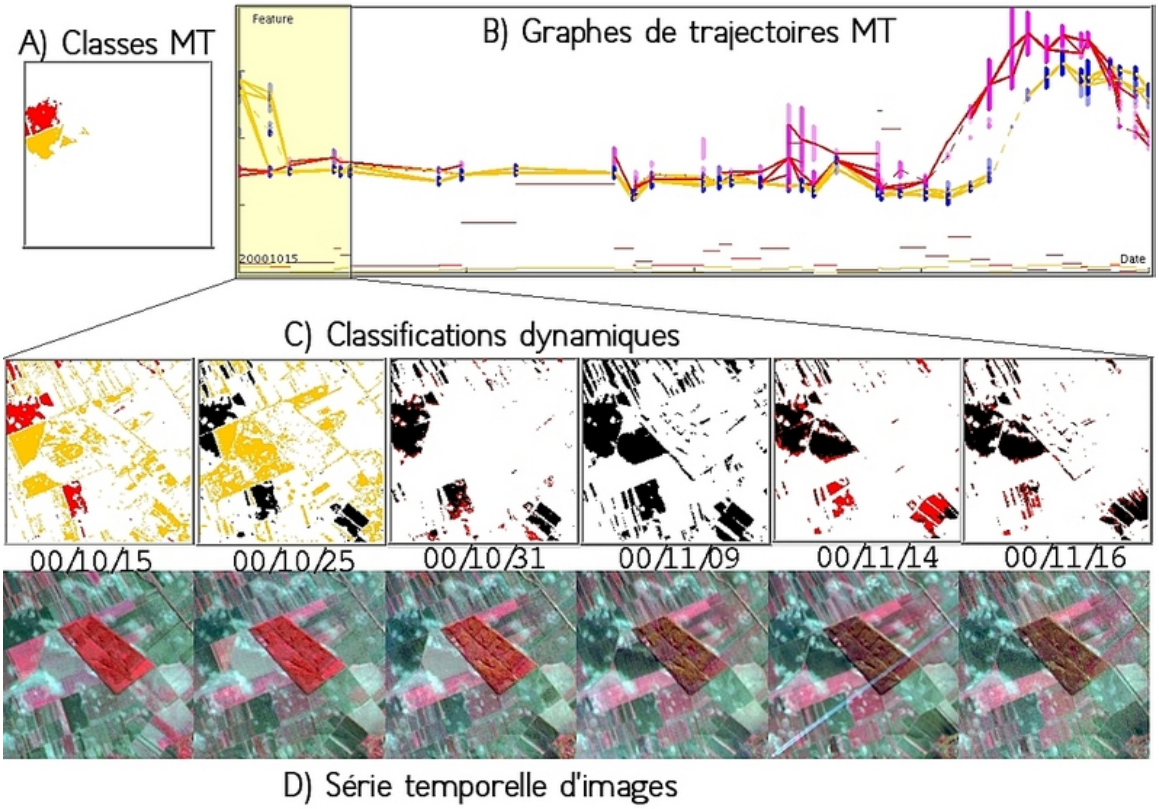


FIG. 6.12 – *Classes MT (A), graphes de trajectoires pour un certain seuil μ (B), classifications dynamiques (C) et série temporelle d'images (D) caractérisant un phénomène de labours progressifs. Chaque classe MT (rouge ou jaune) possède un graphe associé de trajectoires de cluster dynamique (rouge ou jaune). Dans cet exemple, on visualise, sous forme de ligne à hauteur variable (rouge ou jaune), les mesures d'information mutuelle relative à chaque cluster MT (rouge ou jaune) entre deux échantillons temporels successifs. La STIS et les classifications dynamiques présentées dans cet exemple correspondent seulement aux trajectoires dans la période du 15 oct. au 16 nov. 2000 (zone du graphe coloriée en jaune). Les classifications dynamiques sont visualisées dans le même espace spatio-temporel avec l'échelle de couleur suivante : les zones non-communes aux 2 classifications dynamiques possèdent une couleur correspondant à leur classe MT associée (rouge ou jaune) ; les zones communes aux 2 classifications dynamiques sont affichées en noir.*

jaune). Ensuite, la trajectoire se stabilise et est identique à celle d'un champ labouré antérieurement (classe MT rouge).

- Les classifications dynamiques permettent la compréhension de l'évolution spatiale des structures. On remarque que les classifications sont d'abord distinctes et regroupent des champs de culture similaire (jaune) ou des champs labourés (rouge). Au temps suivant, seulement une partie du champ est labourée. Des zones communes (en noir) aux 2 classifications dynamiques sont donc engendrées tandis que les zones non-labourées (en jaune) perdurent. Au troisième échantillon temporel (31 oct. 2000), toutes les zones correspondant aux deux classes MT ont été labourées. On observe donc des classifications dynamiques communes (en noir) présentant l'ensemble des zones labourées de l'image. Les échantillons temporels postérieurs (du 31 oct. au 16 nov. 2000) permettent le suivi des zones labourées.

Comme l'illustre l'analyse de l'exemple précédent, les graphes et les classifications dynamiques constituent une description objective des structures spatio-temporelles présentes dans une STIS. Une fouille visuelle d'information s'appuyant sur ces objets pourra amener à la compréhension de la dynamique de scènes. Toutefois, cette fouille visuelle d'information peut s'avérer laborieuse, d'une part parce que cette analyse visuelle n'est pas intuitive pour un utilisateur non formé, et d'autre part parce que la recherche de motifs similaires peut constituer un travail fastidieux pour l'utilisateur.
En conséquence, en se basant sur cette description objective du contenu informatif des STIS, on développe dans le chapitre suivant un module d'apprentissage interactif des phénomènes intéressants pour l'utilisateur. Par des exemples, définis par l'utilisateur, de structures spatio-temporelles visualisées dans la STIS, dans les graphes ou dans les classifications dynamiques, le système apprendra les labels sémantiques attachés aux différentes structures spatio-temporelles.

6.3 Résumé

Dans ce chapitre, nous avons proposé des méthodes d'apprentissage non-supervisées de structures spatio-temporelles. Ces méthodes ont permis l'inférence de graphes et de classifications dynamiques. L'objectivité de ces descriptions a été mise en évidence par quelques exemples. Pour obtenir ces représentations, les modélisations suivantes ont été développées :

- des modèles stochastiques, appliqués à l'image, ont permis l'extraction de caractéristiques représentées dans les espaces de caractéristiques multitemporel et temporellement localisés ; les modèles de champs aléatoires de Gibbs ont notamment été employés pour extraire des primitives de textures ;

- l'ACP et l'analyse par PP ont été combinées pour la recherche de sous-espaces intéressants dans l'espace de caractéristiques MT ;

- une inférence par le principe MDL a permis l'estimation des paramètres d'un modèle

de mélange de gaussiennes ; des clusters MT et TL ont ainsi été engendrés ;

- finalement, la complémentarité des différentes représentations a été exploitée par un modèle basé sur la divergence de Kullback-Leibler ; cette modélisation a permis l'inférence de trajectoires de clusters dynamiques et, parallèlement, l'inférence de classifications dynamiques.

Chapitre 7

Apprentissage interactif de la sémantique de structures spatio-temporelles

Dans ce paragraphe, on se focalise sur une étape très importante de la fouille d'information : l'interaction avec les utilisateurs et l'incorporation flexible de leurs intérêts spécifiques. Cette étape constitue le dernier niveau de la modélisation hiérarchique de l'information.

7.1 Modélisation bayésienne de sémantiques d'utilisateurs

Les processus spatio-temporels peuvent être indexés par leurs localisations dans l'espace et dans le temps. Ces processus sont susceptibles de posséder des sémantiques subjectives, propres aux utilisateurs. Dans cette perspective, des utilisateurs sont susceptibles d'être intéressés par la recherche d'événements similaires. Ils voudront donc savoir, où et quand, des motifs spatio-temporels similaires ont eu lieu.
Pour répondre à de telles questions, quelques approches ont été expérimentées, en particulier une adaptation du système I2M aux données multitemporelles a été proposée. Cette approche est détaillée dans l'annexe A.2. Mais dans la suite, on développe une approche plus prometteuse qui se base sur le graphe de trajectoires $\mathcal{G}$, qui constitue un codage robuste et non-supervisé des STIS. En effet, les sous-graphes $\mathcal{G}_k$ contenus dans le graphe $\mathcal{G}$ sont des modèles stochastiques pour ces processus spatio-temporels et sont aussi susceptibles de posséder des sémantiques propres aux utilisateurs[1]. En conséquence, basé sur cette caractérisation objective du signal, on s'intéresse maintenant à apprendre les significations subjectives attachées aux graphes $\mathcal{G}_k$, dans le but d'effectuer un étiquetage sémantique des

[1]Contrairement aux notations précédentes (c.f. § 6.2), les graphes $\mathcal{G}_k$ ne représentent pas forcément l'intégralité des trajectoires des clusters MT, mais peuvent aussi représenter seulement des parties d'une trajectoires MT ou même de plusieurs trajectoires MT.

structures spatio-temporelles. Une telle procédure pourra permettre la reconnaissance et la recherche probabiliste d'évènements similaires. La figure 7.1 schématise le dialogue entre l'utilisateur et un système d'apprentissage basé sur les graphes : l'utilisateur transmet au système des fenêtres temporelles et des cartes spatiales correspondant aux structures spatio-temporelles qui l'intéressent ou non ; basé sur les représentations sous forme de graphes associées à ces exemples positifs et négatifs, le système apprend interactivement les sémantiques pour finalement retrouver les localisations temporelles et spatiales d'une collection de structures spatio-temporelles possédant des sémantiques similaires.

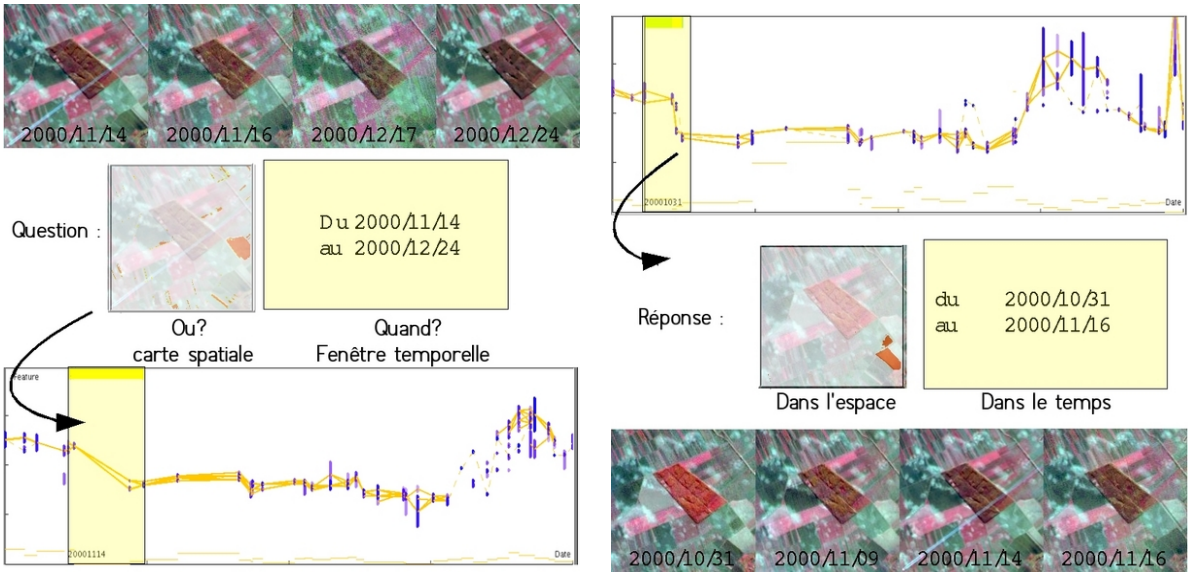


FIG. 7.1 – *Dialogue entre l'utilisateur et un système d'apprentissage basé sur des graphes. L'utilisateur transmet au système des fenêtres temporelles et des masques spatiaux. Ils correspondent à des structures spatio-temporelles d'intérêt ou de non-intérêt. Basé sur une représentation de ces exemples sous forme de graphes et de classifications dynamiques, le système apprend interactivement des sémantiques spécifiques aux utilisateurs, et retrouve dans l'espace et dans le temps des structures spatio-temporelles similaires.*

Dans le but de définir un modèle pour les sémantiques des structures, on développe une mesure paramétrique de similarité $S_\phi(\mathcal{G}_0, \mathcal{G}_k)$ entre deux sous-graphes $\mathcal{G}_0$ et $\mathcal{G}_k$. Cette mesure est une extension de l'algorithme de l'appariement inexact présenté dans la section 4.4.2. Le vecteur de paramètres ϕ pondère les contributions de chaque type d'attribut caractérisant les sous-graphes. Ce modèle paramétrique de similarité sera détaillé dans la section suivante (c.f. § 7.1.1).
Un certain vecteur de paramètres correspond à une similarité particulière. Cette similarité particulière peut formaliser une certaine sémantique d'utilisateur. Il nous faut donc ajuster les paramètres en fonction de la sémantique d'un utilisateur. Nous verrons dans la section 7.1.2, que les paramètres ϕ du modèle de similarité et le graphe de référence $\mathcal{G}_0$ peuvent être appris via un processus d'apprentissage interactif qui se base sur des exemples fournis par un utilisateur. Partant de là, il est possible de relier les éléments subjectifs $\mathcal{A}_\nu$,

représentant les sémantiques d'utilisateur, aux éléments objectifs $\mathcal{G}_k$. Pour cela, on fait l'hypothèse que la similarité paramétrique $S_\phi(\mathcal{G}_0, \mathcal{G}_k)$ constitue un modèle $\mathcal{M}_S$ qui suffit à la description des différentes sémantiques et, on évalue la probabilité de vraisemblance de la sémantique $\mathcal{A}_\nu$ pour chaque graphe $\mathcal{G}_k$. En introduisant la constante de normalisation S_{tot}, on exprime la vraisemblance par

$$p(\mathcal{G}_k \mid \mathcal{A}_\nu, \mathcal{M}_S) = \frac{S_{\widehat{\phi}}(\widehat{\mathcal{G}}_0, \mathcal{G}_k)}{S_{tot}}, \tag{7.1}$$

où $\widehat{\phi}$ et $\widehat{\mathcal{G}}_0$ sont respectivement un vecteur de paramètres et un graphe de référence, tous deux estimés par l'utilisateur. Pour simplifier les notations, on omet dans la suite le conditionnement par le modèle $\mathcal{M}_S$.
Basé sur la vraisemblance, l'utilisation d'un contexte bayésien permet l'inférence des probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$. En effet, en considérant qu'un utilisateur fournit des exemples positifs et négatifs correspondant à une sémantique positive $\mathcal{A}_\nu$ et négative $\neg\mathcal{A}_\nu$, deux probabilités de vraisemblance $p(\mathcal{G}_k \mid \mathcal{A}_\nu)$ et $p(\mathcal{G}_k \mid \neg\mathcal{A}_\nu)$ peuvent être dérivées pour chaque sous-graphe[2]. De ces probabilités, on dérive une distribution a priori sur les graphes par la somme

$$p(\mathcal{G}_k) = \sum_i p(\mathcal{G}_k \mid \mathcal{A}_i)p(\mathcal{A}_i) = p(\mathcal{G}_k \mid \mathcal{A}_\nu)p(\mathcal{A}_\nu) + p(\mathcal{G}_k \mid \neg\mathcal{A}_\nu)p(\neg\mathcal{A}_\nu), \tag{7.2}$$

où la sommation est effectuée sur la sémantique positive et négative. Donc, en employant cette distribution a priori $p(\mathcal{G}_k)$ et une distribution a priori uniforme $p(\mathcal{A}_i)$ sur les sémantiques, les probabilités a posteriori de la sémantique positive peuvent être inférées par

$$p(\mathcal{A}_\nu \mid \mathcal{G}_k) = \frac{p(\mathcal{G}_k \mid \mathcal{A}_\nu)p(\mathcal{A}_\nu)}{p(\mathcal{G}_k)} = \frac{p(\mathcal{G}_k \mid \mathcal{A}_\nu)}{p(\mathcal{G}_k \mid \mathcal{A}_\nu) + p(\mathcal{G}_k \mid \neg\mathcal{A}_\nu)}. \tag{7.3}$$

Par apprentissage interactif des paramètres du modèle de similarités, des probabilités a posteriori de sémantiques spécifiques aux utilisateurs $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ sont obtenues pour chaque sous-graphe $\mathcal{G}_k$. L'inférence de probabilités a posteriori à partir de sémantiques positives et négatives (Eq. 7.3) présente des similarités avec celle développée par *Schroeder et al. (2000)*. Un étiquetage sémantique est produit en attachant une sémantique $\mathcal{A}_\nu$ à un sous-graphe $\mathcal{G}_k$ lorsque la probabilité a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ est supérieure à 0.5. En conséquence, cette procédure amène à un étiquetage sémantique des graphes qui permet la compréhension de scènes dynamiques et la recherche probabiliste de structures spatio-temporelles dans les STIS.

Pour obtenir cet étiquetage sémantique, on doit dans un premier temps définir un modèle paramétrique de similarité $S_{\widehat{\phi}}(\widehat{\mathcal{G}}_0, \mathcal{G}_k)$ entre graphes pour permettre l'évaluation des probabilités $p(\mathcal{G}_k \mid \mathcal{A}_\nu)$. Dans un second temps on doit développer une procédure pour l'apprentissage interactif de la distribution $p(\phi \mid \mathcal{A}_\nu)$ des paramètres du modèle de similarité.

[2]Nous désignons par "sémantique négative" la sémantique définie par l'ensemble des contre-exemples d'utilisateurs. Notons que cette définition de sémantique négative considère un certain ensemble de contre-exemples et non le monde non-clos des contre exemples, et par conséquent cette définition est différente de celle utilisée dans le domaine de la logique.

L'inférence de la distribution des paramètres est suivie d'une étape d'estimation de paramètres et de la mise à jour de la fonction de similarité. Ces points font l'objet des trois sous chapitres suivants.

7.1.1 Modèle paramétrique de similarité entre graphes

Les sous-graphes $\mathcal{G}_k$ sont des caractéristiques temporelles et multidimensionnelles particulières qui représentent des parties de trajectoires de clusters MT. L'information est condensée dans les noeuds et les arcs. Reprenons les quantités définies dans la section 6.2.2 et introduisons quelques notations. Un sous-graphe $\mathcal{G}_k$ associé à un certain cluster MT, est constitué de noeuds qui représentent des clusters TL. On note par $\boldsymbol{\nu_k} = \{\nu_i^k\}$ la collection de noeuds du sous-graphe $\mathcal{G}_k$. Relativement à celui-ci, chaque noeud ν_i^k de la collection est caractérisé par une surface de recouvrement des classes MT et TL que l'on notera $\pi_{\nu_i^k}$, des paramètres gaussiens TL $\xi_{\nu_i^k}$ et une mesure de divergence. De plus, une classe spatiale est associée à ce noeud. Un arc du sous-graphe $\mathcal{G}_k$, représentant l'évolution du cluster MT entre deux échantillons temporels, est caractérisé par un intervalle de temps T, un flux de pixels entre deux clusters TL γ, l'évolution de paramètres gaussiens $\delta(\xi)$ des clusters TL et le changement intra-cluster quantifié par l'information mutuelle MI. Pour l'ensemble des arcs dirigés vers le noeud ν_i^k, on notera les flux et les évolutions des paramètres gaussien par $\gamma_{\nu_i^k}$ et $\delta_{\nu_i^k}(\xi)$

Pour prendre en considération ces différents attributs, on étend l'algorithme de l'appariement inexact (c.f. § 4.4.2), à un modèle paramétrique de similarités entre sous-graphes, où le vecteur de paramètres pondère la contribution des différents attributs de graphe dans la fonction de coût. On s'attache à définir la similarité des sous-graphes $\mathcal{G}_1$ et $\mathcal{G}_2$ composés par les noeuds $\boldsymbol{\nu_1} = \{\nu_i^1\}$ et $\boldsymbol{\nu_2} = \{\nu_i^2\}$. On emploie une collection supplémentaire de noeuds $\lambda = \{\lambda_i\}$ pour définir une transformation globale

$$\mathcal{F} = \{f\}, \tag{7.4}$$

composée par une certaine combinaison de transformations élémentaires

$$f : \nu^1 \rightarrow \nu^2 \cup \lambda = \nu^{2\lambda}, \tag{7.5}$$

ou chaque transformation associe un noeud ν_i^1 de $\mathcal{G}_1$, à un noeud de ν_i^2 de $\mathcal{G}_2$ ou bien à un noeud supplémentaire λ_i. Une fonction de coût $C_\phi(f(\nu_i^1) = \nu_j^{2\lambda})$ est associée à chaque transformation élémentaire. Dans un soucis de simplicité on choisit de définir cette fonction de coût sous la forme d'une somme pondérée de similarité relatives aux différents attributs des graphes. Un vecteur de paramètres $\phi = \{\phi_l\}$ pondère donc chaque élément de la somme. La fonction de coût $C_\phi(f(\nu_i^1) = \nu_j^{2\lambda})$ s'exprime donc par

$$\phi_1\Delta(\pi_{\nu_i^1}, \pi_{\nu_j^{2\lambda}}) + \phi_2\Delta(\xi_{\nu_i^1}, \xi_{\nu_j^{2\lambda}}) + \phi_3\Delta(\gamma_{\nu_i^1}, \gamma_{\nu_j^{2\lambda}}) + \phi_4\Delta(\delta_{\nu_i^1}(\xi), \delta_{\nu_j^{2\lambda}}(\xi)) \tag{7.6}$$

où $\Delta(.)$ représente une mesure de similarité qui est, soit une différence pour des scalaires, soit une mesure de similarité telle que la divergence de Kullbach-Leibler pour les distributions. Parce qu'un intervalle temporel entre deux échantillons et l'information mutuelle

caractérisent communément tous les arcs d'un certain graphe dans le même intervalle de temps, les similarités entre ces attributs ne doivent être reportées qu'une seule fois par intervalle de temps. De même, lorsqu'on évalue la similarité entre des graphes composés chacun de 2 trajectoires MT, on introduit une nouvelle mesure de similarité dans la somme. Cette mesure évalue la différence de distances, notée $\delta_{MT}(\xi)$, entre les distributions des 2 clusters MT du même sous-graphe. Dans ce cas, on ne reporte cette mesure dans la fonction de coût qu'une seule fois à chaque échantillon temporel. En conséquence, on ajoute à la fonction de coût (Eq. 7.6) une seule fois par intervalle de temps les termes suivants

$$\phi_5 \Delta(MI) + \phi_6 \Delta(T), \tag{7.7}$$

et lorsqu'on considère des graphes composés par 2 trajectoires MT, on ajoute à la fonction de coût (Eq. 7.6) une seule fois par échantillon temporel le terme $\phi_7 \Delta(\delta_{MT}(\xi))$.
On précise que lorsque l'on associe un noeud ν_i^1 de $\mathcal{G}_1$ avec un noeud supplémentaire λ_i, les similarités avec les attributs de $\mathcal{G}_1$ sont calculées par rapport à la valeur 0 pour les scalaires, et pour les distributions par rapport à la distribution du cluster MT associée à $\mathcal{G}_2$ et projetée au temps considéré.
La fonction de coût globale n'est autre que la somme des fonctions de coût élémentaires $C_\phi(f(\nu_i^1) = \nu_j^{2\lambda})$ sur tous les noeuds de ν_i^1. Finalement, la mesure de similarité est ensuite définie pour un certain paramètre ϕ, comme la combinaison de transformations élémentaires, associée à la fonction de coût globale minimale

$$S_\phi(\mathcal{G}_1, \mathcal{G}_2) = \min_{\mathcal{F}} \big(\sum_i C_\phi(f(\nu_i^1) = \nu_j^{2\lambda}) \big). \tag{7.8}$$

On note par ailleurs que parce que l'équation 7.6 est une simple somme, on peut décomposer le coût global $C_\phi(f(\nu_i^1) = \nu_j^{2\lambda})$ en une somme pondérée de coûts partiels relatifs aux différents attributs de graphes dans la mesure de similarité. Ces coûts partiels, notés $C_l(f(\nu_i^1) = \nu_j^{2\lambda})$, sont donc pondérés par les différents paramètres ϕ_l du vecteur ϕ. En conséquence, peut écrire

$$C_\phi(f(\nu_i^1) = \nu_j^{2\lambda}) = \sum_{l=1}^{7} \phi_l C_l(f(\nu_i^1) = \nu_j^{2\lambda}). \tag{7.9}$$

En notant par $S_l(\mathcal{G}_1, \mathcal{G}_2)$ le coût relatif au paramètre ϕ_l dans la fonction de similarité $S_\phi(\mathcal{G}_1, \mathcal{G}_2)$, l'Eq. 7.8 devient

$$S_\Phi(\mathcal{G}_1, \mathcal{G}_2) = \sum_{l=1}^{7} \phi_l \min_{\mathcal{F}} \big(\sum_i C_l(f(\nu_i^1) = \nu_j^{2\lambda}) \big) = \sum_{l=1}^{7} \phi_l S_l(\mathcal{G}_1, \mathcal{G}_2). \tag{7.10}$$

Le vecteur de paramètres $\phi = \{\phi_l\}$ pondère donc une somme de contributions $S_l(\mathcal{G}_1, \mathcal{G}_2)$ relatives aux différents attributs de graphes dans la mesure de similarité.

Pour parvenir à la mesure de similarité, l'on doit donc rechercher le coût global minimum induit par une association particulière de noeuds et d'arcs. La recherche s'effectue sur toutes les combinaisons possibles entre les éléments des deux sous-graphes. L'historique d'un tel processus de recherche peut être enregistré dans une structure appelée arbre

de recherche. La solution est représentée par une ramification, partant du tronc vers une des branches, associée à un coût global minimum. Mais, comme précisé déjà dans la section 4.5.1, souvent la solution est difficilement accessible car il existe une explosion combinatoire de solutions ou en d'autres termes, une explosion de ramifications à évaluer.
Pour faire face au coût calculatoire trop important que nécessitent ces évaluations, nous utilisons une solution simple consistant à développer seulement les ramifications de l'arbre les plus intéressantes, tout au long de sa construction. Plus précisément, on place d'abord une limite de nombre de noeuds de l'arbre à considérer durant la recherche. On développe ensuite toutes les ramifications de l'arbre à partir du tronc, jusqu'à ce que cette limite soit dépassée. Alors, on procède à l'élagage des ramifications de l'arbre les plus coûteuses, afin de se replacer sous la limite et l'on continue le développement des ramifications restantes de l'arbre. On réitère la procédure jusqu'à la constitution du noeud le plus haut de l'arbre. La solution est alors constituée par la ramification la moins coûteuse. Notons que l'algorithme peut toutefois converger vers des minima locaux et que cette approche est sous-optimale. Ce problème d'explosion combinatoire de solutions pourra être plus efficacement résolu en considérant des approches comme par exemple la programmation dynamique.

7.1.2 Apprentissage interactif de la distribution des paramètres

Dans la section précédente, nous avons développé une mesure de similarité entre graphes de trajectoires, dépendante d'un vecteur paramètre ϕ. Ce vecteur de paramètres pondère les contributions relatives aux différents attributs de graphes dans la fonction de similarité. Or, nous avons vu qu'un certain vecteur de paramètres correspond à une similarité particulière qui elle même peut formaliser une certaine sémantique relative à un utilisateur. Mais, l'ajustement manuel des paramètres du vecteur afin de définir une mesure de similarité spécifique à la sémantique de l'utilisateur est une tâche fastidieuse. En conséquence, on emploie une procédure d'apprentissage supervisé pour obtenir une estimation $\widehat{\phi}$ du vecteur de paramètres ϕ. Cet estimateur, utilisé dans l'équation 7.1, permettra l'évaluation des probabilités de vraisemblance $p(\mathcal{G}_k \mid \mathcal{A}_\nu)$ et $p(\mathcal{G}_k \mid \neg\mathcal{A}_\nu)$, qui par l'équation 7.3 amènent aux probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$.
Dans la suite, nous nous concentrons sur l'obtention de la vraisemblance de la sémantique positive. Toutefois, on applique la même procédure pour obtenir la vraisemblance de la sémantique négative. On présentera d'abord un modèle multinomial pour l'approximation de la distribution continue du vecteur de paramètres ϕ, puis on introduira une procédure pour l'apprentissage interactif de ce modèle.

7.1.2.1 Modèle multinomial pour des paramètres discrétisés

L'idée pour l'estimation interactive des paramètres du modèle de similarité en fonction d'une certaine sémantique est la suivante : on considère un certain graphe de référence $\mathcal{G}_0$ et un exemple, fourni par l'utilisateur, de graphe $\mathcal{G}_k$ possédant une certaine sémantique $\mathcal{A}_\nu$; alors, plus le coût $S_l(\mathcal{G}_0, \mathcal{G}_k)$ relatif au paramètre ϕ_l est faible, plus la pondération du paramètre ϕ_l est susceptible d'être importante. En d'autres termes, on fait l'hypothèse que chaque fonction de coût $S_l(\mathcal{G}_0, \mathcal{G}_k)$, relative au paramètre ϕ_l, est proportionnelle à l'opposé

de la valeur du paramètre ϕ_l

$$\phi_l \propto -S_l(\mathcal{G}_0, \mathcal{G}_k). \tag{7.11}$$

On exploite maintenant cette hypothèse de proportionnalité. D'abord, pour rendre comparables entre eux les différents paramètres ϕ_l, on normalise au préalable les intervalles où les fonctions $S_l(\mathcal{G}_0, \mathcal{G}_k)$ prennent leurs valeurs. Ensuite, comme l'estimation d'une distribution continue est difficile lorsque l'on dispose de très peu de données, on discrétise les paramètres continus $\{\phi_l\}$ en r niveaux de quantification, pour que chaque paramètre ϕ_l prenne ses valeurs dans $\{\phi_l^1, ..., \phi_l^r\}$ et suive une distribution multinomiale. Cette distribution présente l'avantage de posséder des paramètres correspondant à des probabilités physiques qui, comme nous le verrons, pourront être estimées en temps réel dans un cadre bayésien.

Le nombre de niveaux r doit être suffisamment grand pour que la distribution discrète approxime la distribution continue. [3] Donc, en considérant une sémantique propre à un utilisateur $\mathcal{A}_\nu$, la distribution de probabilités conditionnelles de chaque paramètre ϕ_l est définie par

$$p(\phi_l = \phi_l^j \mid \omega, \mathcal{A}_\nu) = p(\Lambda(S_l(\mathcal{G}_0, \mathcal{G}_k)) = \phi_l^j \mid \omega, \mathcal{A}_\nu) = \omega_j, j = 1, ..., r \tag{7.12}$$

où $\omega = \{\omega_1, ..., \omega_r\}$ sont les paramètres du modèle multinomial[4] et $\Lambda(.)$ est un opérateur qui discrétise l'intervalle normalisé où les fonctions $S_l(\mathcal{G}_0, \mathcal{G}_k)$ prennent leurs valeurs, en r niveaux de quantification $\{\phi_l^1, ..., \phi_l^r\}$. Pour simplifier les notations, $p(\phi_l = \phi_l^j \mid \omega, \mathcal{A}_\nu)$ sera noté $p(\phi_l = \phi_l^j \mid \mathcal{A}_\nu)$.
De plus, on fait l'hypothèse d'indépendance statistique entre les paramètres de telle sorte que l'estimation de la distribution de probabilités conjointe ne soit pas nécessaire. Notons que cette hypothèse est nécessaire pour réduire la complexité du modèle et ainsi permettre l'apprentissage interactif présenté dans la suite. Cependant, la validité de cette hypothèse dépend de la nature des termes $S_l(\mathcal{G}_0, \mathcal{G}_k)$ que pondèrent les paramètres ϕ_l considérés dans le modèle de similarité. Par exemple, le terme relatif au flux de points sera vraisemblablement indépendant du terme relatif à la distribution des clusters. Par contre, le terme relatif à l'information mutuelle ne sera pas forcément indépendant du terme relatif au flux de points. En supposant cette hypothèse valide, on obtient donc

$$p(\phi \mid \mathcal{A}_\nu) = p(\phi_1 \mid \mathcal{A}_\nu) p(\phi_2 \mid \mathcal{A}_\nu) ... \tag{7.13}$$

7.1.2.2 Apprentissage interactif bayésien

Pour un certain jeu d'exemples de sémantique $\mathcal{A}_\nu$, on déplace maintenant la discussion de l'évaluation des distribution de probabilités $p(\phi_j \mid \mathcal{A}_\nu)$ pour chaque paramètre ϕ_j à l'évaluation de la distribution des probabilités $p(\omega \mid \xi)$ des paramètres ω d'un modèle multinomial.

[3] L'ordre de grandeur du nombre r doit être fonction du nombre d'exemples fournis par l'utilisateur durant le processus d'apprentissage. Dans ces travaux, on prend un nombre r égal à 1000.

[4] Notons que le paramètre ω_1 est donné par $1 - \sum_{j=2}^{r} \omega_j$.

L'apprentissage interactif développé dans cette section est inspiré des travaux de *Heckerman (1999)* et de *Schroeder (2000)* sur l'apprentissage avec des réseaux bayésiens.
L'estimation est effectuée par entraînement d'un réseau bayésien par un utilisateur. Une approche bayésienne est adoptée du fait de sa robustesse lorsque seulement peu d'exemples d'utilisateur sont disponibles. L'utilisateur fourni un jeu noté T d'exemples de sous-graphes possédant la sémantique désirée. Avec ces exemples d'utilisateurs, on définit pour chaque paramètre ϕ_l, un vecteur $N = \{N_1, ..., N_r\}$ où N_j désigne le nombre d'occurrences de ϕ_l^j, c'est à dire le nombre de fois que $\phi_l = \phi_l^j$ apparaît dans les exemples de T. Notons que les paramètres ω du modèle multinomial (Eq. 7.12) correspondent à des probabilités d'occurrence.

Pour l'évaluation interactive de ces probabilités d'occurrence (ou des paramètres du modèle multinomial), on introduit comme a priori conjugué (c.f. § 4.7) la distribution de Dirichlet. Pour un certain stade de connaissance ξ, cette distribution dépend d'un vecteur d'hyper-paramètres $\alpha = \{\alpha_1, ..., \alpha_r\}$ et s'exprime par

$$p(\omega \mid \xi) \;=\; Dir(\omega \mid \alpha_1, ..., \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{j=1}^{r} (\omega_j)^{\alpha_j - 1} \tag{7.14}$$

où $\alpha = \sum_{j=1}^{r} \alpha_j$ et $\alpha_j > 0, \forall j \in [1, r]$ et où $\Gamma(x)$ désigne la fonction Gamma.
Pour initialiser l'apprentissage, on utilise une distribution a priori conjugué définie par la distribution de Dirichlet où tous les hyper-paramètres $\alpha^{(0)} = \{\alpha_1^{(0)}, ..., \alpha_r^{(0)}\}$ sont fixés à 1. Cette distribution a priori est une distribution uniforme sur ω ; Elle s'écrit

$$p(\omega) = Dir(\omega \mid \alpha_1^{(0)}, ..., \alpha_r^{(0)});\ \forall j \in [1, r], \alpha_j^{(0)} = 1. \tag{7.15}$$

Après que l'utilisateur ait fourni un jeu d'exemples $T^{(1)}$ correspondant à un vecteur d'occurrence $\{N_j^{(1)}\}$, en utilisant la règle de Bayes, on dérive la distribution a posteriori

$$\begin{aligned} p(\omega \mid T^{(1)}) &= \frac{p(T^{(1)} \mid \omega) p(\omega)}{p(T^{(1)})} \\ &= Dir(\omega \mid \alpha_1^{(0)} + N_1^{(1)}, ..., \alpha_r^{(0)} + N_r^{(1)}) \\ &= Dir(\omega \mid \alpha_1^{(1)}, ..., \alpha_r^{(1)}) \end{aligned} \tag{7.16}$$

Puis, quand de nouveau l'utilisateur fourni un nouveau jeu d'exemples $T^{(2)}$ qui est supposé indépendant de $T^{(1)}$ et qui correspond à un nouveau vecteur d'occurrence $\{N_j^{(2)}\}$, on obtient une nouvelle distribution a posteriori

$$\begin{aligned} p(\omega \mid T^{(2)}, T^{(1)}) &= \frac{p(T^{(2)} \mid \omega, T^{(1)}) p(\omega \mid T^{(1)})}{p(T^{(2)})} \\ &= Dir(\omega \mid \alpha_1^{(1)} + N_1^{(2)}, ..., \alpha_r^{(1)} + N_r^{(2)}) \end{aligned} \tag{7.17}$$

où les nouveaux hyper-paramètres sont calculés en additionnant chaque ancien hyper-paramètre, avec le nombres de fois que $\phi_l = \phi_l^j$ apparaît dans le jeu d'exemples $T^{(2)}$.
Ainsi, chaque nouveau jeu d'exemples $T^{(i)}$ fourni par l'utilisateur peut servir à la mise à

jour de nouveaux hyper-paramètres : $\alpha_j^{(i)} = \alpha_j^{(i-1)} + N_j^{(i)}$.

Pour un certain stade d'apprentissage issu de la mise à jour du vecteur d'hyper-paramètres α avec un ensemble d'exemples T, l'estimation de $p(\phi_l^j \mid \mathcal{A}_\nu, T)$ est obtenue en utilisant l'estimateur MMSE (c.f. § 4.1.4) pour le paramètre ω_j :

$$p(\phi_l = \phi_l^j \mid \mathcal{A}_\nu) = E[\omega_j] = \int \omega_j p(\omega \mid T) d\omega = \frac{\alpha_j}{\alpha}. \tag{7.18}$$

Finalement, avec l'équation 7.13, on obtient $p(\phi \mid \mathcal{A}_\nu)$ en effectuant le produit $\prod_l p(\phi_l \mid \mathcal{A}_\nu)$.

7.1.3 Estimation et mise à jour de la similarité

Un certain stade d'apprentissage permet l'inférence de la distribution de probabilités $p(\phi \mid \mathcal{A}_\nu)$. On peut alors utiliser l'estimateur MMSE pour estimer le vecteur de paramètres ϕ de la fonction de similarité

$$\widehat{\phi}_{MMSE} = E[\phi], \tag{7.19}$$

où $E[.]$ désigne l'espérance relative à la distribution $p(\phi \mid \mathcal{A}_\nu)$. On note que la distribution multinomiale n'exhibe pas un maximum clair du fait du peu d'exemples fournis par l'utilisateur comparé au grand nombre r de valeurs ϕ_l^j; $l = 1, ..., r$. Ceci justifie donc que l'estimateur MMSE soit plus adapté à cette situation que l'estimateur MAP. Cette mise à jour du vecteur de paramètres permet la mise à jour de la fonction de similarité $S_{\widehat{\Phi}_{MMSE}}(\mathcal{G}_0, \mathcal{G}_k)$. Ainsi, par l'équation 7.1, les probabilités de vraisemblance de sémantique $\mathcal{A}_\nu$ peuvent être réévaluées pour chaque sous-graphe $\mathcal{G}_k$

$$p(\mathcal{G}_k \mid \mathcal{A}_\nu) = \frac{S_{\widehat{\phi}_{MMSE}}(\widehat{\mathcal{G}}_0, \mathcal{G}_k)}{S_{tot}}, \tag{7.20}$$

où S_{tot} est un facteur de normalisation. Notons que pour initialiser le vecteur de paramètres de la fonction de similarité, on emploie une distribution uniforme sur les ϕ_l.
On remarque par ailleurs que l'arbre de recherche issu d'une recherche de similarité avec un certain graphe de référence $\widehat{\mathcal{G}}_0$ et avec un certain vecteur $\widehat{\phi}_{MMSE}$, peut être dans certains cas réutilisé à l'itération suivante. En effet, pour calculer la fonction de similarité lorsque l'on observe des petites variations du vecteur de paramètres entre deux itérations et que l'on utilise le même graphe de référence, on peut utiliser un arbre de recherche de morphologie identique. Cet arbre de morphologie identique, doit toutefois voir les poids de ses noeuds être mis à jour en fonction des variations du vecteur de paramètres. Pour évaluer la nouvelle fonction de similarité, il est seulement nécessaire de réévaluer le chemin le moins coûteux allant du tronc vers la plus haute ramification de cet arbre. Cependant, lorsque le graphe de référence est réestimé, il faut alors reformer intégralement l'arbre de recherche. La réutilisation des arbres de recherche permet de réduire la complexité calculatoire du processus d'apprentissage interactif qui, de par sa définition, se doit d'être

effectué en temps réel.

L'inférence des probabilités de vraisemblance permet, par l'équation 7.3, l'évaluation des probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ pour chaque graphe $\mathcal{G}_k$. On obtient alors un nouveau graphe de référence estimé $\widehat{\mathcal{G}_0}$ en sélectionnant le graphe $\mathcal{G}_k$ possédant la probabilité a posteriori maximum

$$\widehat{\mathcal{G}_0} = arg \max_{\mathcal{G}_k} p(\mathcal{A}_\nu \mid \mathcal{G}_k). \tag{7.21}$$

L'actualisation du graphe de référence sera prise en compte à l'itération suivante.

7.1.4 Attribution de labels sémantiques

La modélisation bayésienne de sémantiques a permis l'évaluation de probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ pour chaque graphe $\mathcal{G}_k$ à partir d'exemples d'utilisateur. On se base sur ces probabilités pour attribuer aux graphes, i.e. aux structures spatio-temporelles, des labels sémantiques. On considère qu'une structure possède une sémantique $\mathcal{A}_\nu$ si la probabilité a posteriori associée est supérieure à un certain seuil. Ce seuil dépend de la probabilité de fausse détection désirée par l'utilisateur. Ce seuil doit toutefois être supérieur à une valeur de 0.5. Une valeur supérieure à 0.5 signifie que les structures labellisées ont plus de chance d'être associées à la sémantique $\mathcal{A}_\nu$ que de n'y être pas associée.

Le processus bayésien d'apprentissage interactif présenté ci-dessus a permis l'étiquetage sémantique des structures spatio-temporelles composant la STIS, par l'intermédiaire des sous-graphes. Les labels sémantiques ont été attribués aux structures, avec plus ou moins de certitude, en fonction de probabilités a posteriori.

7.2 Indicateurs de pertinence

L'évaluation de la qualité des labels sémantiques attribués aux structures spatio-temporelles permet à l'utilisateur de juger de la qualité de l'apprentissage par le système. On présente dans cette section différents critères servant à cette évaluation *(Héas et al., 2004b)*.

Des critères peuvent être obtenus de l'analyse directe de la distribution a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$.
D'abord, il est évident que pour une certaine sémantique $\mathcal{A}_\nu$, la maximisation sur $\mathcal{G}_k$ des probabilités a posteriori aboutissant à une valeur proche de l'unité est révélateur d'un apprentissage sémantique concluant.
Ensuite, la moyenne et la variance peuvent être des critères de qualité plus globaux. Pour une collection de p structures spatio-temporelles différentes, la moyenne statistique de la

distribution a posteriori est définie par

$$E\{p(\mathcal{A}_\nu \mid \mathcal{G}_k)\} = \frac{1}{p}\sum_{k=1}^{p} p(\mathcal{A}_\nu \mid \mathcal{G}_k), \tag{7.22}$$

et sa variance est définie par

$$\delta^2 p(\mathcal{A}_\nu \mid \mathcal{G}_k) = \frac{1}{p}\sum_{k=1}^{p} \Big(p(\mathcal{A}_\nu \mid \mathcal{G}_k) - E\{p(\mathcal{A}_\nu \mid \mathcal{G}_k)\}\Big)^2. \tag{7.23}$$

La variance de la distribution a posteriori nous renseigne globalement sur le niveau d'homogénéité des structures relativement à une certaine sémantique. Par exemple, une faible variance correspond à une collection de structures homogènes, c'est à dire de sémantiques similaires, tandis qu'une forte variance indique que les structures de la collection sont hétérogènes, c'est à dire fortement discriminées par la sémantique considérée. La moyenne, quant à elle, nous renseigne sur le niveau global d'attachement de la collection de structures à une certaine sémantique.

Des indicateurs de pertinence peuvent être obtenus par l'analyse des distributions du vecteur de paramètres de similarité relatifs à une sémantique positive $p(\phi \mid \mathcal{A}_\nu)$ et une sémantique négative $p(\phi \mid \neg\mathcal{A}_\nu)$.
D'abord la convergence du processus d'estimation peut s'évaluer en mesurant les variations de la norme du vecteur de paramètres, qui a été estimé pour chaque différente sémantique entre deux itérations de la procédure d'apprentissage. Cette variation entre deux itérations (i) et $(i+1)$ est égale à

$$C = \|\widehat{\phi}_{MMSE}^{(i+1)} - \widehat{\phi}_{MMSE}^{(i)}\|, \tag{7.24}$$

où $\|.\|$ représente la norme euclidienne.
Ensuite, il est aussi intéressant de quantifier le niveau de différentiation entre la sémantique positive et la sémantique négative. En effet, plus les deux sémantiques seront dissociées, plus les labels sémantiques seront pertinents. Pour accéder à cet indicateur de pertinence, on emploie la divergence symétrisée de Kullback-Leibler. La divergence $D_{sym}\Big(p(\phi \mid \mathcal{A}_\nu), p(\phi \mid \neg\mathcal{A}_\nu)\Big)$ entre les distributions de paramètres relatives aux sémantiques positives et négatives renseigne donc l'utilisateur sur le niveau global de différenciation de la sémantique positive et de la sémantique négative. On reprend l'hypothèse d'indépendance entre les paramètres effectuée précédemment pour décomposer cette mesure de différentiation en une somme de mesures de divergence

$$D_{sym}\Big(p(\phi \mid \mathcal{A}_\nu), p(\phi \mid \neg\mathcal{A}_\nu)\Big) = \sum_{l=1}^{7} D_{sym}\Big(p(\phi_l \mid \mathcal{A}_\nu), p(\phi_l \mid \neg\mathcal{A}_\nu)\Big). \tag{7.25}$$

Le l-ème élément de la somme de l'équation 7.25 est représentatif de la distance entre les distributions uni-dimensionnelles $p(\phi_l \mid \mathcal{A}_\nu)$ et $p(\phi_l \mid \neg\mathcal{A}_\nu)$, qui respectivement sont

relatives à une sémantique positive et négative.[5] Donc, chaque élément de la somme est un indicateur de la qualité des labels sémantiques relativement à un type d'attributs de graphes que pondère le paramètre ϕ_l. Ainsi par ces mesures, on peut quantifier le degré de contribution d'un certain attribut de graphes dans la constitution du label sémantique. Ainsi, on pourra par exemple évaluer l'importance du "flux de points caractéristiques" dans l'élaboration d'un label sémantique "occlusions".

Durant la procédure d'apprentissage, ces indicateurs de pertinences sont évalués en temps-réel.

7.3 Un exemple d'apprentissage interactif

On présente maintenant l'apprentissage interactif d'une sémantique de maturation de champs. On effectue l'étiquetage sémantique d'une portion de la série temporelle ADAM. Cette portion est définie par une fenêtre spatiale de 200x200 pixels et une fenêtre temporelle constituée de 24 images acquises du 13 mars au 26 août 2001. On décompose le graphe de trajectoires en K_{MT} collections de sous-graphes correspondant aux différents clusters MT. Chaque collection comprend tous les sous-graphes possibles, constitués de 12 échantillons temporels et inclus dans la STIS de 24 images.
Des exemples de structures spatio-temporelles sont alors transmis au système. La figure 7.2 présente 2 exemples positifs et 2 exemples négatifs relatifs à la maturation de champs choisis pour cet apprentissage. Après chaque exemple de l'utilisateur, le système réévalue les probabilités des sous-graphes et répond à la requête de l'utilisateur en lui présentant une collection de structures spatio-temporelles possédant les plus importantes probabilités a posteriori. La figure 7.2 présente les 7 structures spatio-temporelles, définies dans une classe spatiale et dans une fenêtre temporelle, possédant les probabilités les plus fortes.
Par ailleurs, pour aider l'utilisateur dans sa recherche, le système réévalue les indicateurs de pertinence après chaque itération du processus d'apprentissage. Ces indicateurs sont présentés à l'utilisateur pour le guider dans le choix de ses exemples et pour lui permettre d'évaluer la qualité courante de son étiquetage sémantique. Ils sont représentés dans la figure 7.3.

Examinons de plus près le processus d'apprentissage. Le premier exemple positif fourni par l'utilisateur entraîne un premier étiquetage sémantique pertinent. Cependant, les structures les plus probables ne correspondent pas à des probabilités très élevées. De plus, beaucoup de structures possèdent des probabilités proches du maximum. Un deuxième exemple positif similaire au premier est fourni au système. Pour les structures les plus probables, ce nouvel exemple ne cause pas de grandes variations. Mais par contre pour

[5] Parce que l'on effectue le calcul de la divergence entre des distributions multinomiales évaluées interactivement avec peu d'exemples d'utilisateur, la distribution peut être fortement discontinue. Pour une meilleure évaluation de la divergence de l'équation 7.25, on procède d'abord à la convolution de la distribution multinomiale avec un noyau gaussien possédant une variance fonction du nombre d'exemples fournis par l'utilisateur *(Antoniadis, 2001)*.

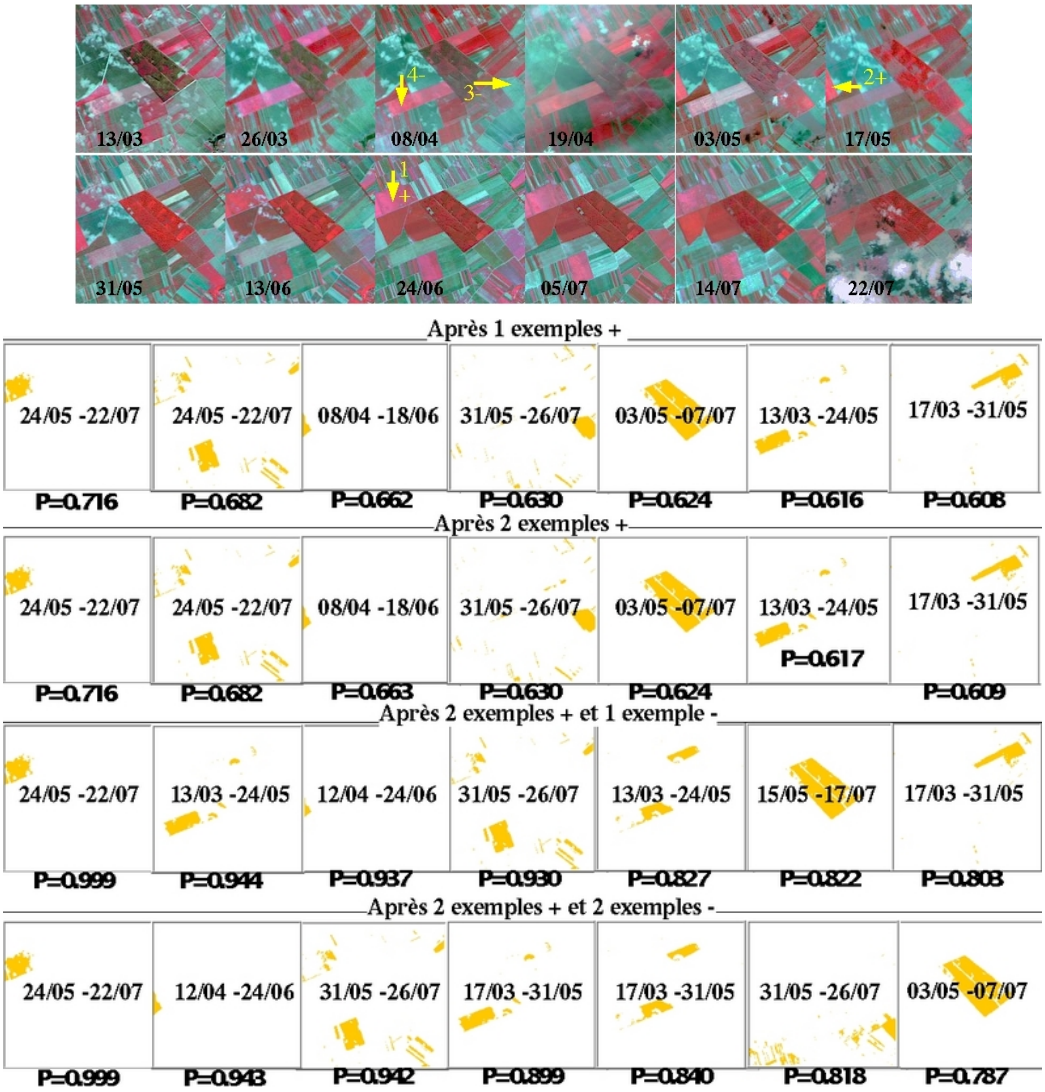


FIG. 7.2 – *Apprentissage interactif d'une sémantique de maturation de champs. (haut) La série temporelle est représentées par 12 des 24 images utilisées pour l'apprentissage. Deux exemples positifs puis 2 exemples négatifs sont donnés dans l'ordre croissant du temps. Ces exemples sont présentés par l'intermédiaire des flèches jaunes. Ces dernières indiquent un certain champ au temps correspondant à l'apogée de sa maturation. (bas) Collections de structures spatio-temporelles de probabilités a posteriori P maximum retrouvée après chaque exemple de l'utilisateur donné successivement dans le temps. Chaque ligne représente la collection courante de structures spatio-temporelles retrouvée qui possédent les probabilités les plus importantes. Ces structures sont définies par des classes spatiales affichées en jaune et par des fenêtres temporelles indiquées au centre des classes.*

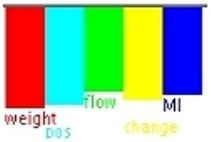
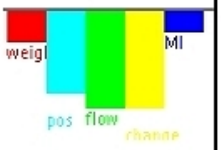
	Apprentissage 1	Apprentissage 2	Apprentissage 3	Apprentissage 4
Exemple	+	+	-	-
Moyenne	0.4791	0.4785	0.4959	0.4951
Variance	0.1147	0.1168	0.1814	0.1741
C+	2,2361	0.2930	0.0044	0.0293
C-	0,00000	0,0000	1,5492	0.4212
Divergence				

FIG. 7.3 – *Evolution des indicateurs de pertinence avec le degré d'apprentissage. Après chaque exemple d'utilisateur le tableau présente de haut en bas : la moyenne $E\{p(\mathcal{A}_\nu \mid \mathcal{G}_k)\}$ et la variance $\delta^2 p(\mathcal{A}_\nu \mid \mathcal{G}_k)$ des probabilités a posteriori ; les variations de la norme C du vecteur de paramètres pour la sémantique positive $+$ et négative $-$; les histogrammes de divergences relatives aux niveaux de différentiation entre la sémantique positive et négative pour chaque attribut (en rouge π, en cyan ξ, en vert γ, en jaune $\delta(\xi)$ et en bleu MI).*

des structures à la sémantique encore mal discriminée (pas présentées dans la figure 7.2), des variations plus importantes de probabilités sont observées. Afin de véritablement discriminer la sémantique des structures, deux exemples négatifs sont communiqués au système. Certaines structures se voient alors attribuées des probabilités proches de l'unité, alors que d'autres voient leurs probabilités chuter.

Les constatations précédentes sont confortées par l'analyse des indicateurs de pertinences (c.f. Fig. 7.3). D'abord, on peut juger de la qualité de l'apprentissage en observant l'évolution de la moyenne statistique des probabilités a posteriori. En effet, on observe que la moyenne a tendance à diminuer avec le degré d'apprentissage. Comme les phénomènes de maturation de champs ne sont pas majoritaires dans la base de données, cette diminution témoigne de la qualité de l'apprentissage.
Ensuite, on observe que la variance des probabilités a posteriori a, quant à elle, tendance à augmenter. Le premier exemple négatif cause notamment une augmentation non-négligeable. Cette augmentation témoigne d'une discrimination croissante entre les structures de sémantique positive et celles de sémantique négative.
L'analyse de l'évolution des critères $C+$ et $C-$, montre des variations brutales de la norme des vecteurs de paramètres après le premier exemple positif et négatif. Les variations causées par le quatrième exemple et par d'autres exemples non-présentés ici, ont tendance à être moins brutales. Ces faible variations indiquent que les estimations successives des vecteurs de paramètres convergent avec l'augmentation du degré d'apprentissage.
Enfin, les évolutions des histogrammes de divergence présentent des changements brutaux et il est difficile de tirer des conclusions à aprtir de ces 4 exemples. En continuant à communiquer de nouveaux exemples au système (non-présentés ici), les variations des histogrammes s'attenuent considérablement. Plusieurs exemples sont donc nécessaires pour

clairement différencier la sémantique négative de la sémantique positive. On peut conclure de l'observation du dernier histogramme que trois attributs de graphes (la distribution des clusters TL ξ, le flux de points γ et les variations de distributions $\delta(\xi)$) contribuent principalement à l'élaboration du modèle sémantique.

7.4 Résumé

Dans ce chapitre, nous avons proposé une méthodologie originale pour l'apprentissage interactif de sémantiques de structures spatio-temporelles composant les STIS. Des sémantiques propres aux utilisateurs sont modélisées par l'intermédiaire d'exemples positifs et négatifs. L'apprentissage permet l'attribution probabiliste de labels sémantiques aux sous-graphes de trajectoires inférées précédemment.
Cet étiquetage est obtenue par inférence bayésienne. On introduit d'abord un modèle paramétrique de similarité entre graphes. Ensuite, on formule l'hypothèse qu'une sémantique particulière correspond à un certain jeu de paramètres du modèle de similarité. Basé sur cette hypothèse on effectue l'estimation interactive des paramètres du modèle via un apprentissage interactif bayésien : l'intégration progressive des exemples d'utilisateur est effectuée en appliquant un modèle de Dirichlet dans un réseau bayésien.
La pertinence de l'approche est finalement mise en évidence par un exemple d'apprentissage interactif.

Troisième partie

Système de fouille visuelle d’information

Chapitre 8

Description et évaluation du système de fouille visuelle d'information

Dans ce chapitre, nous décrivons et évaluons le système de fouille. Nous présentons d'abord l'architecture client-serveur et les différentes composantes du système. Nous nous attachons par ailleurs à expliquer le dialogue entre l'utilisateur et le système par l'intermédiaire de l'interface Homme-Machine. Nous achevons la description du système en abordant, d'un point de vue plus technique, la chaîne de traitement. Nous évaluons ensuite l'apprentissage du système avec la base de données ADAM : la qualité des différentes étapes de la modélisation est évaluée et des exemples de compréhension dynamique de scènes sont présentés.

8.1 Le système de fouille visuelle d'information

8.1.1 Architecture client-serveur du système de fouille

L'architecture du système de fouille I2M *(Datcu et al., 2003)*, qui est présenté dans l'annexe A.1, est globalement préservée et étendue au cas spatio-temporel. Après quelques adaptations, on parvient donc à l'architecture client-serveur présentée dans la figure 8.1. Du coté du serveur, toutes les données de l'archive de STIS subissent dans un premier temps une phase d'ingestion nécessitant des calculs intensifs. Par cette phase d'ingestion de données, on génère de façon automatique à partir des données, une collection de graphes et de classifications dynamiques.
Une fois les données ingérées, du coté du client, un utilisateur a d'abord la possibilité de visualiser les données. Pour cela, par l'intermédiaire d'une interface Homme-Machine, l'utilisateur accède à diverses formes de représentations des structures spatio-temporelles. L'utilisateur peut visualiser des sous-graphes de trajectoires de clusters dynamiques, qui ont tous été au préalablement transmis au client. Conjointement à la représentation sous forme de graphes, l'utilisateur peut aussi visualiser des morceaux de STIS dans leurs représenta-

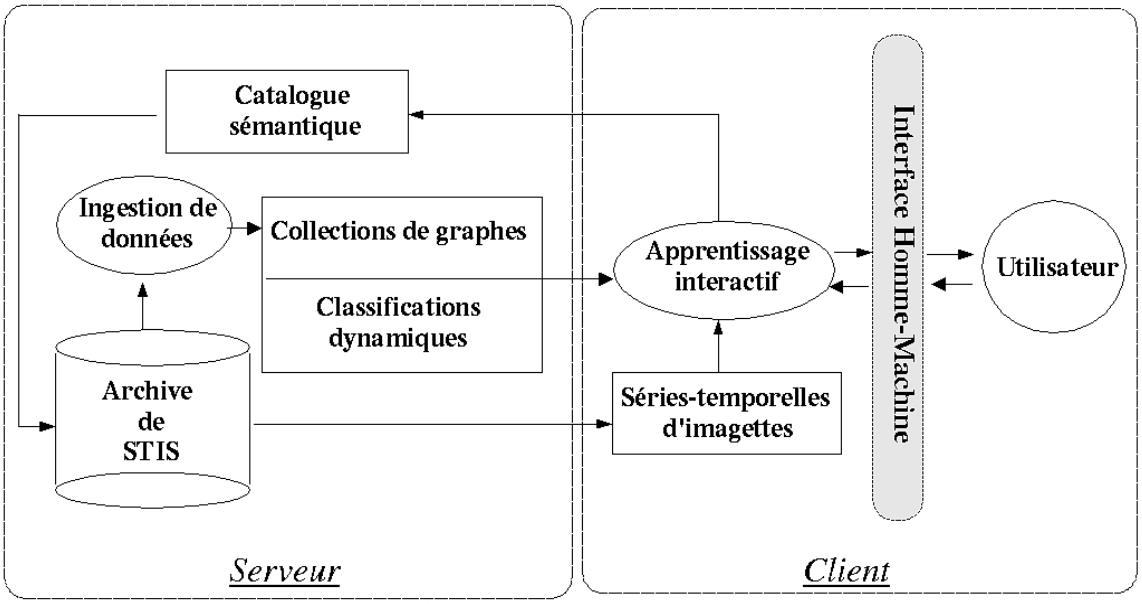


FIG. 8.1 – *Schématisation de l'architecture client-serveur du système de fouille de séries temporelles d'images satellites.*

tions naturelles par l'intermédiaire de séries temporelles d'imagettes. Celles-ci possèdent des radiométries modifiées. Ces imagettes sont créées "à la volée" à partir de l'archive et communiquées sur demande au client. En rapport avec les graphes, des classifications dynamiques sont aussi visualisables par l'utilisateur. De même, des parties de classifications dynamiques sont transmises sur demande au client.
Initialement, la fouille se concentre sur une seule STIS choisie aléatoirement dans la base de données. La visualisation est initialisée aléatoirement avec une des structures appartenant à cette sous-base de données. L'utilisateur a ensuite la possibilité de naviguer, dans l'espace et le temps, au travers des différentes représentations pour choisir des exemples positifs et négatifs de structures spatio-temporelles. Cette navigation implique des requêtes du client vers le serveur. En effet, pour mettre à jour les visualisations, des parties de classifications dynamiques et des séries temporelles d'imagettes doivent être transmises au client.
Le processus d'apprentissage interactif débute lorsque, pour une certaine sémantique, l'utilisateur communique au système la localisation spatio-temporelle (masque spatial et fenêtre temporelle) correspondant à une première structure d'intérêt. Cette structure d'intérêt, i.e. sous-graphe d'intérêt, est sélectionnée parmi une collection constituée par l'ensemble des structures, i.e. sous-graphes, de la STIS considérée. Alors, pour chaque exemple correspondant à un certain sous-graphe de trajectoires, le système met à jour les probabilités a posteriori associées à chacun des sous-graphes de la collection, ainsi que les indices de pertinence. Le coût calculatoire nécessaire pour cette mise à jour est relativement faible et l'apprentissage est effectué interactivement. Les structures spatio-temporelles associées aux sous-graphes sont alors rangées en fonction de leurs probabilités. Une collection ordonnée de structures est ainsi formée. Les localisations spatio-temporelles des éléments les plus probables de la collection sont ensuite transmises au serveur, pour que celui-ci re-

tourne au client les données nécessaires à la visualisation des structures les plus probables. Cette visualisation renseigne qualitativement l'utilisateur sur la pertinence de l'apprentissage. Pour accéder à une évaluation plus quantitative, l'utilisateur peut visualiser le niveau des indicateurs de pertinence. Si l'utilisateur est insatisfait de l'étiquetage sémantique, de nouveaux exemples positifs et négatifs peuvent être fournis au système. Pour améliorer la définition du label sémantique, l'utilisateur a la possibilité de continuer l'apprentissage avec une autre STIS de la base de données. Ce processus d'apprentissage est itéré jusqu'à satisfaction de l'utilisateur.
Si l'étiquetage sémantique est jugé satisfaisant par l'utilisateur, l'utilisateur peut rechercher des structures similaires dans l'archive de STIS entière. Cette recherche probabiliste se base sur la maximisation des probabilités a posteriori $p(\mathcal{A}_\nu \mid \mathcal{G}_k)$, qui dépend seulement des paramètres du modèle de similarité $\mathcal{G}_0$ et ϕ. En conséquence, pour retrouver la collection ordonnée de structures, seuls ces paramètres doivent être sauvegardés dans le catalogue sémantique.
Pour mettre à jour l'archive avec de nouvelles données, la phase d'ingestion doit seulement être appliquée sur les nouvelles STIS de l'archive. Par l'utilisation du catalogue sémantique, l'étiquetage sémantique des nouvelles composantes de l'archive est effectué automatiquement sans aucun nouveau processus d'apprentissage. Toutefois, une opportunité pour parfaire l'apprentissage et faire évoluer le catalogue sémantique est toujours laissée à l'utilisateur.

8.1.2 Interface Homme-Machine

Par l'interface Homme-Machine, un dialogue nécessaire à l'apprentissage est rendu possible entre l'utilisateur et le système. Du coté de l'utilisateur, des exemples sont sélectionnés au travers de diverses représentations visuelles, et du coté du système, des étiquetages sémantiques sont retournés à l'utilisateur en association avec des mesures de qualité. L'interface supporte donc la visualisation de structures particulières dans diverses représentations, la navigation dans l'archive, la visualisation de labels sémantiques ainsi que la visualisation des indicateurs de pertinence.

L'utilisateur visualise des séries temporelles d'imagettes dans l'espace de représentation naturelle des STIS. Pour chaque requête de l'utilisateur, on forme à la volée une série temporelle d'imagettes correspondant à une certaine portion d'une STIS de l'archive. Des standards de visualisation (e.g. JPEG ou MPEG-4) sont utilisés. Les imagettes affichées présentent une scène sous-échantillonnée. L'utilisateur accède aussi en pleine résolution à des fenêtres spatiales de la scène. La navigation temporelle et spatiale s'effectue respectivement par l'incrémentation de l'index temporel des images et par la sélection d'une zone géographique de la STIS.
Pour atteindre une meilleure homogénéité visuelle entre images de la série, on fait l'hypothèse que pour chaque bande spectrale de chaque image de la STIS, les réflectances recherchées possèdent une même distribution Normale. Cette hypothèse permet de construire, par l'intermédiaire d'une adaptation dynamique des couleurs, une série temporelle d'imagettes aux réflectances normalisées. Cette modification d'histogramme s'obtient pour chaque image en deux étapes : dans un premier temps, on égalise les images,

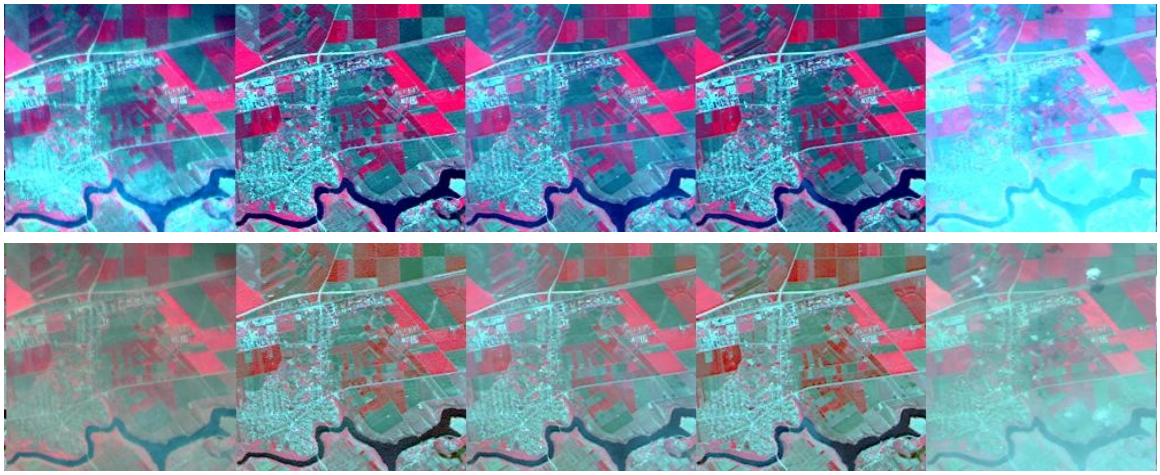


FIG. 8.2 – *Série temporelle d'images visualisée avec une table de couleur linéaire (en haut) et gaussienne (en bas).*

en utilisant pour fonction de transfert l'inverse de la fonction de répartition des réflectances de l'image ; puis dans un deuxième temps on transforme la distribution uniforme des histogrammes en une loi gaussienne de moyenne et de variance fixée.
La série temporelle d'imagettes sera visuellement plus intuitive car, même si l'homogénéité crée est artificielle, l'analyse visuelle sera moins perturbée par les discontinuités temporelles des radiométries. Ces discontinuités, atténuées par l'ajustement linéaire des radiométries (c.f. § 2.2.3), sont pour la plupart dues à la présence d'aérosols. La figure 8.2 présente le résultat du processus de normalisation. On remarque toutefois que, globalement, les couleurs des STIS sont plus "fades" après normalisation. De plus, par ces modifications d'histogrammes, on atténue l'évolution globale des radiométries.

Les sous-graphes de trajectoires de clusters dynamiques sont représentés dans l'espace des caractéristiques évolutives. L'utilisateur sélectionne une composante du graphe multidimensionel et visualise les trajectoires associées à un ou plusieurs clusters dynamiques. La complexité des graphes de trajectoires peut être ajustée par l'utilisateur en modifiant le seuil μ. Les différents attributs des graphes sont visualisés par des segments aux couleurs nuancées. On visualise par exemple la surface de recouvrement des classes MT, la moyenne et la variance des clusters par des segments verticaux possédant respectivement une couleur plus ou moins saturée, une certaine position et une certaine longueur.
L'utilisateur peut naviguer dans le temps par simples clics sur les localisations temporelles d'intérêt. La sélection d'une nouvelle localisation temporelle implique simultanément la mise à jour de l'index temporel des imagettes à visualiser. D'autre part, il navigue dans l'espace en sélectionnant des classes MT associées aux clusters MT directement dans la STIS. La sélection des classes est cependant plus aisée au travers des classifications dynamiques.

L'utilisateur accède à la visualisation des classifications dynamiques dans l'espace spatio-

temporel. Après voir sélectionné une ou deux classes spatiales (ou clusters MT) et une fenêtre temporelle d'intérêt, l'utilisateur peut observer l'évolution spatiale des structures. Cette visualisation est effectuée en employant un maximum de trois couleurs différentes. Ces dernières désignent les classes TL associées à chacune des deux classes MT et les classes TL communes aux deux classes MT. La navigation spatio-temporelle s'effectue en correspondance avec les représentations précédentes. D'autre part, en ajustant le paramètre de seuil μ, les classifications dynamiques engloberont plus ou moins de classes TL.

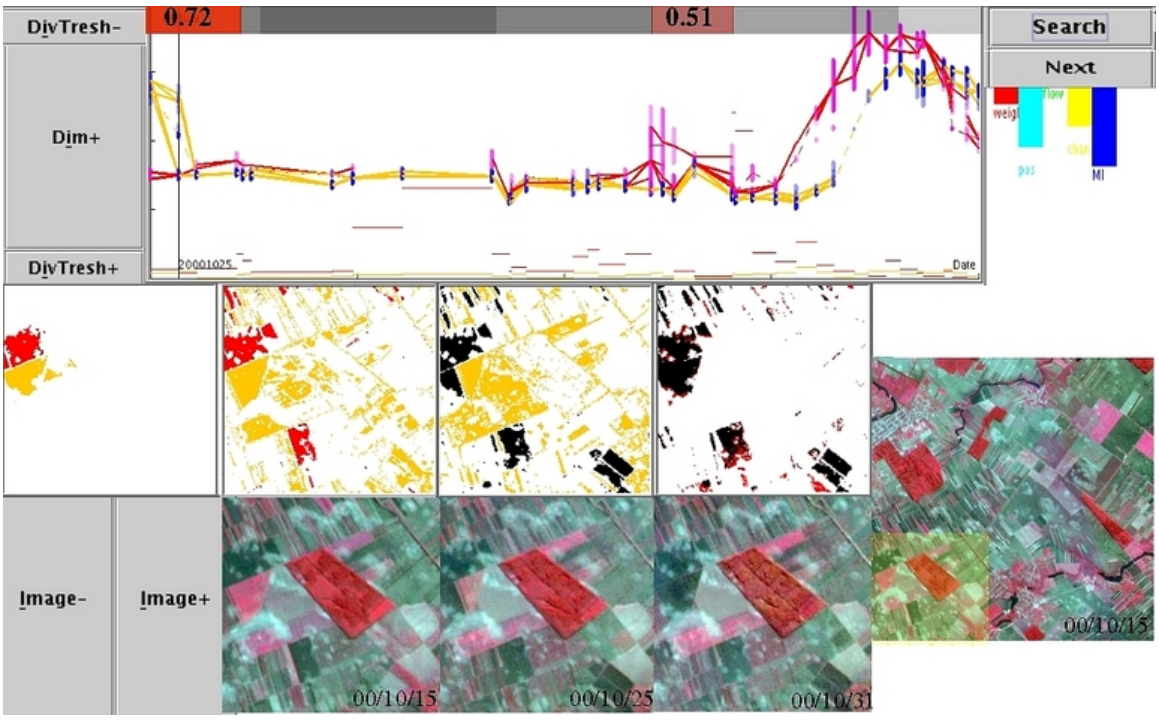


FIG. 8.3 – *Maquette de l'interface graphique intégrée dans le système de fouille.*

L'utilisateur formule les exemples de structures par l'intermédiaire de fenêtres temporelles et de classes MT. Ces paramètres sont communiqués au système soit en sélectionnant les structures dans les représentations précédentes, soit en remplissant directement un champ "texte".
Pour un certain degré d'apprentissage, l'utilisateur peut visualiser l'étiquetage sémantique correspondant. Des labels rouges sont attachés aux sous-graphes possédant des probabilités a posteriori supérieure à 0.5. La couleur rouge de ces labels est nuancée proportionnellement aux probabilités a posteriori. Les éléments de la collection de structures spatio-temporelles sont ordonnés en fonction des probabilités a posteriori. Pour juger de la qualité de l'étiquetage, l'utilisateur visualise les structures les plus probables ou les moins probables. L'analyse de chaque structure de la collection implique la mise à jour des représentations. L'évaluation de la qualité de l'apprentissage est aussi visualisée au travers des barres de divergences et par l'affichage des mesures de convergence et d'homogénéité. Après chaque itération du processus d'apprentissage, les labels et les différentes mesures de pertinence sont mises à jour.

8.1.3 Implémentation de la chaîne de traitements

Dans la suite, nous présentons la chaîne de traitement nécessaire à l'insertion d'une nouvelle STIS dans le système. Afin de simplifier les traitements postérieurs, on découpe la scène globale en sous-scènes de taille réduite. Les sous-scènes dynamiques formées sont sauvegardées sur un espace disque du serveur avec leur localisation géographique et leur temps d'acquisition. Différentes caractéristiques sont ensuite extraites pour chaque série temporelle. Pour chaque type de caractéristique extraite et pour chaque série temporelle, un graphe est généré. Ceux-ci sont sauvegardés sur le disque avec leurs classifications MT et TL correspondantes. Les traitements ci-dessus sont effectués hors-ligne et les résultats sont sauvegardés sur le serveur. Ils sont codés en C/C++ et impliquent un coût calculatoire important.
Pour un certain type de caractéristique, l'utilisateur définit une sémantique de structures en employant un sous-graphe contenu dans un des graphes. Il peut à tout moment étendre sa recherche aux autres graphes de la base de données.
L'interface Homme-Machine décrite dans la section précédente commande la procédure d'apprentissage interactif. L'interface est une applet codée en JAVA accessible par le client via Internet. Les algorithmes d'apprentissage interactif, qui sont des applets et servlets JAVA, doivent aussi être lancés sur la machine client.

Malheureusement, par manque de temps, les algorithmes implémentés ne possèdent pas toutes les fonctionnalités précédemment décrites. D'abord, l'extension d'un étiquetage sémantique aux autres graphes de la base de données n'a pas pu être implémentée. A l'heure actuelle, le prototype disponible se limite donc à la fouille d'un seul graphe à la fois. Cependant, on peut étendre manuellement la recherche dans l'archive entière en introduisant comme estimation a priori, les paramètres de similarités estimés précédemment.
Ensuite, l'implémentation des algorithmes d'apprentissage a été effectuée en C++ et non en JAVA. Le prototype disponible à l'heure actuel n'est donc pas accessible en ligne. Pour permettre la communication entre l'interface Homme-Machine codée en JAVA et l'algorithme d'apprentissage codé en C++, des fichiers "texte" sont employés. L'interface Homme-Machine sert donc pour l'instant à la visualisation des données ainsi que des résultats de l'apprentissage (labels sémantiques, indicateurs de pertinence). Enfin, à défaut de spécifier les exemples en sélectionnant les zones spatio-temporelles d'intérêt directement dans les différents espaces de représentation, les localisations des exemples de structures spatio-temporelles sont entrées manuellement par l'utilisateur pour alimenter la procédure d'apprentissage.

8.2 Evaluation du système de fouille

8.2.1 Evaluation de l'apprentissage non-supervisé de scènes dynamiques ADAM

Une évaluation exhaustive nécessite l'analyse des performances de tous les modules du système de fouille avec des données variées. Nous nous attacherons à effectuer des

analyses, du moins qualitativement, pour diverses scènes dynamiques de la STIS ADAM.

En accord avec le schéma de modélisation hiérarchique bayésienne du contenu informatif des STIS, on commence par l'analyse du contenu informatif des données et par l'évaluation de la qualité de l'extraction de caractéristiques. Pour décrire le contenu des STIS ont utilise des caractéristiques spectrales et texturales des différentes images de la STIS.
Le système développé dans cette thèse possède une procédure d'extraction de caractéristiques comparables au système I2M développé par *Schroeder (2000)*. Pour ce système, l'évaluation de la qualité de l'extraction de caractéristiques spectrales et de texture a été effectuée par *Datcu et Seidel (2002)*. Nous référons donc le lecteur a ce document pour l'analyse détaillée de ce module. Celui-ci admet qu'un bruit additionnel gaussien de variance inconnue détériore les images. Il applique donc un filtre moyenneur pour déterminer le rapport signal-bruit relatif à chaque bande spectrale. Il quantifie ainsi la quantité d'information spectrale disponible dans chaque bande spectrale. Pour la texture modélisée par les champs aléatoires de Gibbs, il utilise la borne de Rao-Cramer pour évaluer la qualité du vecteur de paramètres estimés.

Procédons maintenant à l'évaluation des performances du processus de réduction de dimensionnalité. Pour cela, concentrons nous sur la réduction de dimensionnalité de caractéristiques spectrales. On reprend l'analyse par PP de l'espace MT des caractéristiques entrepris dans la section 6.1.2.2. Cet espace MT est formé à partir des 20 premières composantes principales qui représentent plus de 90 pourcent de l'énergie du signal. La figure 8.4 présente les indices de projection des 10 premières composantes extraites.
D'abord, en rapport avec les travaux menés par *Bachmann et al (1995)*, on observe que les composantes extraites contiennent majoritairement des perturbations. En effet, les 3 premières projections bidimensionnelles extraites présentent des nuages. L'extraction de projections présentant ces objets est sans grand intérêt. Le processus de réduction de dimensionnalité est donc détérioré par ceux-ci car l'information sur ces objets est extraite au détriment d'information plus pertinente. En conséquence, un meilleur processus de sélection des images sans nuages (c.f. § 2.2.1) pourra sans doute conduire à une meilleure analyse par PP.
Ensuite, pour l'analyse de cet espace MT (de dimension égale à 20, comprenant 800x800 points caractéristiques), l'indice de la 10-ième projection est environ 10 fois supérieur à l'indice de projection théorique limite calculé pour une P-Valeurs de 0.01. La suppression des images présentant des nuages contribuera bien sûr à la diminution de cet écart. Néanmoins, il subsistera souvent un nombre trop important de composantes indépendantes dans l'espace MT des caractéristiques. On procède donc à la sélection des $l < q$ composantes indépendantes les plus structurées. Le nombre l est ici fixé d'une manière heuristique en s'appuyant sur l'évolution de la valeur de l'indice de projection en fonction des composantes. Des recherches sont donc encore nécessaires pour une détermination automatique du nombre l.

Nous continuons l'évaluation du système par l'analyse de la qualité des collections de clusters et classes engendrées par modélisation des distributions TL et MT. Les performances

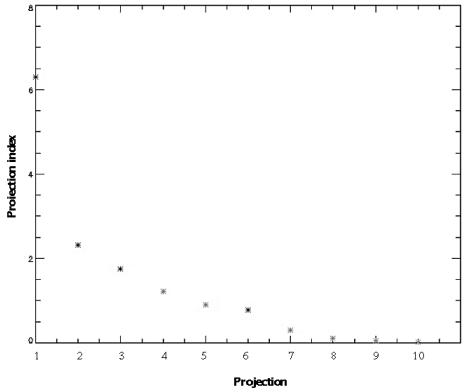


FIG. 8.4 – *Indices de projections associés à l'extraction de l'espace MT des caractéristiques par PP, des 10 premières composantes bidimensionnelles.*

de l'algorithme de modélisation par mélange de gaussiennes basé sur le principe MDL ont été en parties évaluées qualitativement dans la section 6.1.3.2 pour des jeux de données synthétiques. On poursuit par une évaluation quantitative de la modélisation avec les données ADAM.

Comparons dans un premier temps la qualité des classifications MT engendrées par les modélisations d'espaces MT avec une classification de référence faite à partir de données acquises sur le terrain pour le projet ADAM. Pour effectuer la comparaison, on se sert de 3 classifications MT. Un des espaces MT considérés pour la classification est formé à partir des 6 premières composantes principales. Celles-ci représentent plus de 70 pourcent de l'énergie du signal. Les deux autres espaces MT considérés sont formés à partir de composantes indépendantes : un espace regroupe les 3 projections bidimensionnelles les plus structurées et l'autre regroupe les 3 premières projections bidimensionnelles dénuées de perturbations atmosphériques obtenues par PP. La figure 8.5 présente la classification de référence avec sa légende et les 3 classifications MT obtenues. Les classes MT sont évidement plus nombreuses que les classes de références. En effet, cette classification de référence ne différencie pas les différentes évolutions des zones de même nature (e.g. un champ de blé moissonné en juin ne se différencie pas d'un champ de blé moissonné en juillet). De plus, la classification de référence faite au sol ne prend pas en compte, par nature, les effets atmosphériques causes de différenciations de classes. Pour pouvoir effectuer une comparaison visuelle, en fonction des classes de référence, plusieurs classes MT ont été regroupées et représentées avec une même couleur. On remarque clairement que la classification MT la plus proche de la classification de référence est celle n'utilisant que les composantes principales. En effet, pour celle-ci la majorité des champs sont classés identiquement. Toutefois, il existe certaines différences. Les classifications MT succédant à une analyse par PP s'éloignent plus de celle de références. La présence de nuages est vraisemblablement la cause des erreurs et imprécisions de classifications observées. Même dans le cas où les projections bi-dimensionnelles présentant des nuages sont écartées, on observe la discrimination de certaines classes, absentes dans la classifications MT précédente. Néanmoins, il subsiste toujours des imprécisions de classification. Elles sont

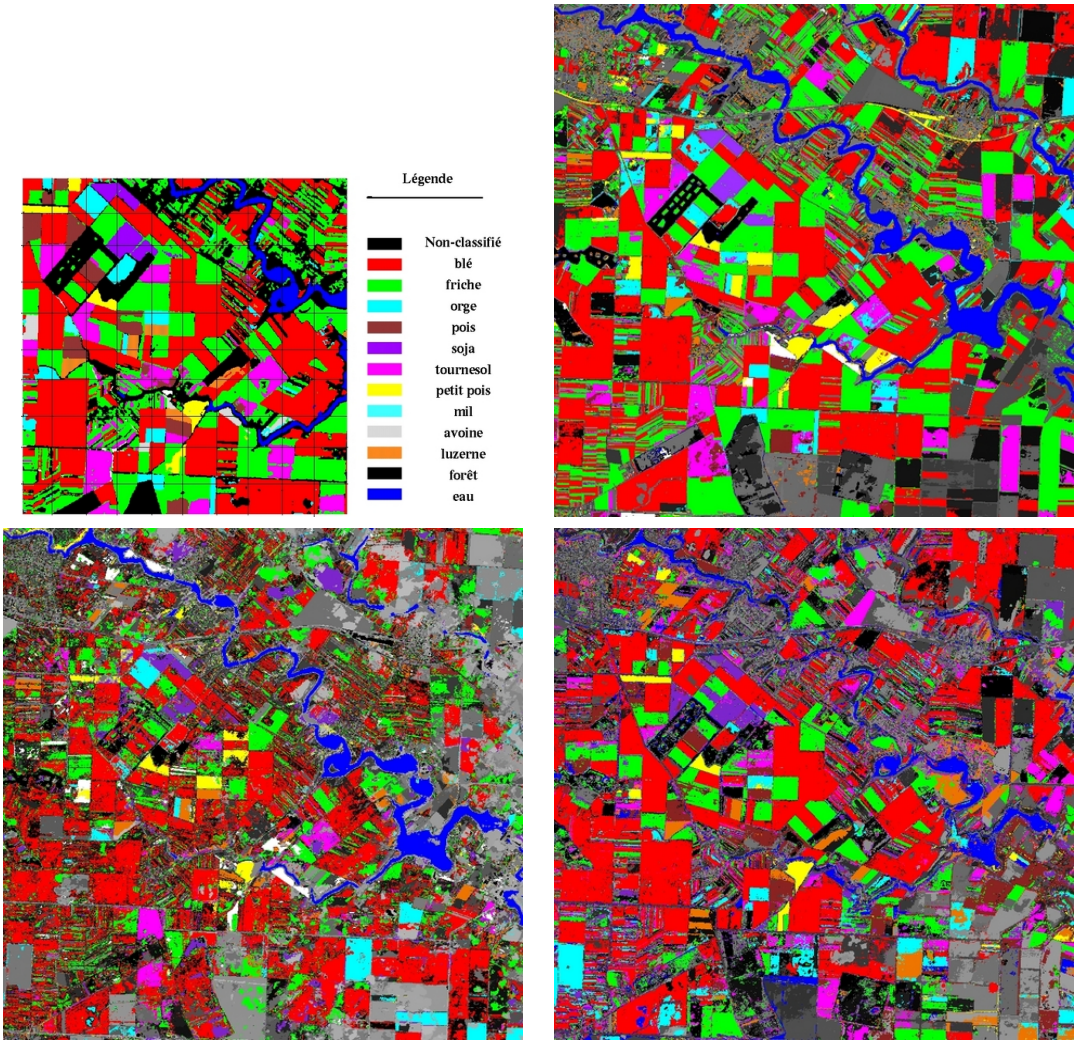


FIG. 8.5 – *Classification de référence légendée (en haut à gauche) ; classifications MT colorisées en fonction de la légende de référence : suite à une ACP (en haut à droite), suite à une analyse par PP (en bas à gauche) et suite à une analyse par PP où les projections présentant une perturbation nuageuse ont été supprimées (en bas à droite).*

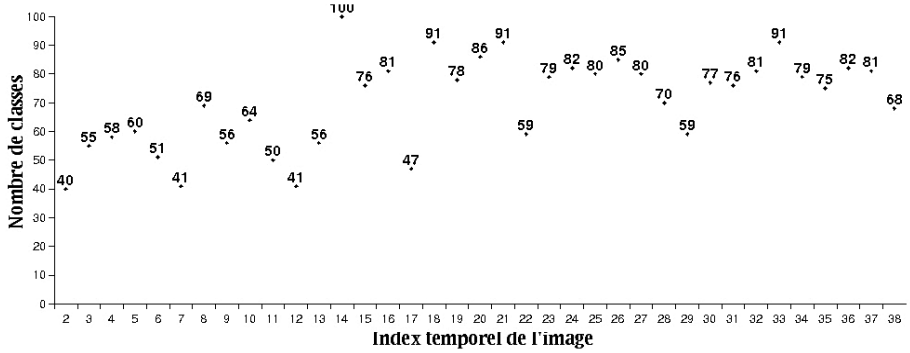


FIG. 8.6 – *Evolution temporelle du nombre de classes de caractéristique spectrale. L'axe des abscisses représente l'index des 38 images en ordre chronologique et l'axe des ordonnées représente le nombre de classes.*

vraisemblablement causées par la suppression d'une information pertinente présente dans les projections nuageuses non considérées lors du processus de classification.
Analysons dans un deuxième temps les classifications TL. En modélisant pour chacune des 38 images de la STIS, les caractéristiques spectrales TL prises dans une fenêtre spatiale de 800x800 pixel, on obtient une série de collections de clusters. L'évolution temporelle du nombre de clusters est présentée dans la figure 8.6. On trouve que les collections possèdent en moyenne un nombre de 70 clusters avec un écart type de ce nombre égale à 15,35. Le nombre moyen de clusters est cohérent avec les résultats d'évaluations précédentes *(Schroeder, 2000)*. On remarque qu'il existe des tendances mais que parfois ce nombre varie brutalement. Ces ruptures correspondent souvent à des événements dans la STIS comme l'apparition d'aérosols ou le passage de nuages. En effet, en observant individuellement chaque classification, on note généralement que le nombre de clusters trouvé est cohérent avec l'inspection visuelle de l'image correspondante. Certaines de ces ruptures ne peuvent pas s'expliquer par une inspection visuelle de l'image, qui ne peut pas objectivement juger de la complexité d'une distribution.
Néanmoins, en étudiant de plus près les différents clusters dans l'espace des caractéristiques, on note que la complexité du modèle de mélange est dans certains cas surestimée. On peut en partie attribuer cette surestimation du nombre de clusters, à l'unique itération de l'algorithme EM lors de la simulation de la suppression d'une gaussienne (c.f. 6.1.3.2). En effet, lorsque l'on s'approche d'un des minima de longueur de description, plusieurs itérations sont parfois nécessaires pour décider si une gaussienne doit être supprimée du mélange. Une telle modification de l'algorithme, bien que facilement réalisable, n'a pas aboutie faute de temps. Par ailleurs, la procédure d'optimisation ne peut pas garantir l'obtention d'un minimum global. Pour lutter contre l'obtention de minima locaux très différents, une distribution a priori sur le nombre de gaussiennes est considérée dans les algorithmes AutoClass ou Snob. Dans notre cas, cet a priori peut être introduit facilement dans la longueur de description L_0. Enfin, l'hypothèse du modèle de mélange utilisée dans ces expériences, où l'on néglige les recouvrements partiels des gaussiennes, peut parfois être restrictive pour discriminer les composantes de certaines distributions. Précisons aussi que la suppression des approximations de certains termes

de la longueur de description (c.f. 6.1.3.2) pourrait contribuer à l'amélioration des résultats.

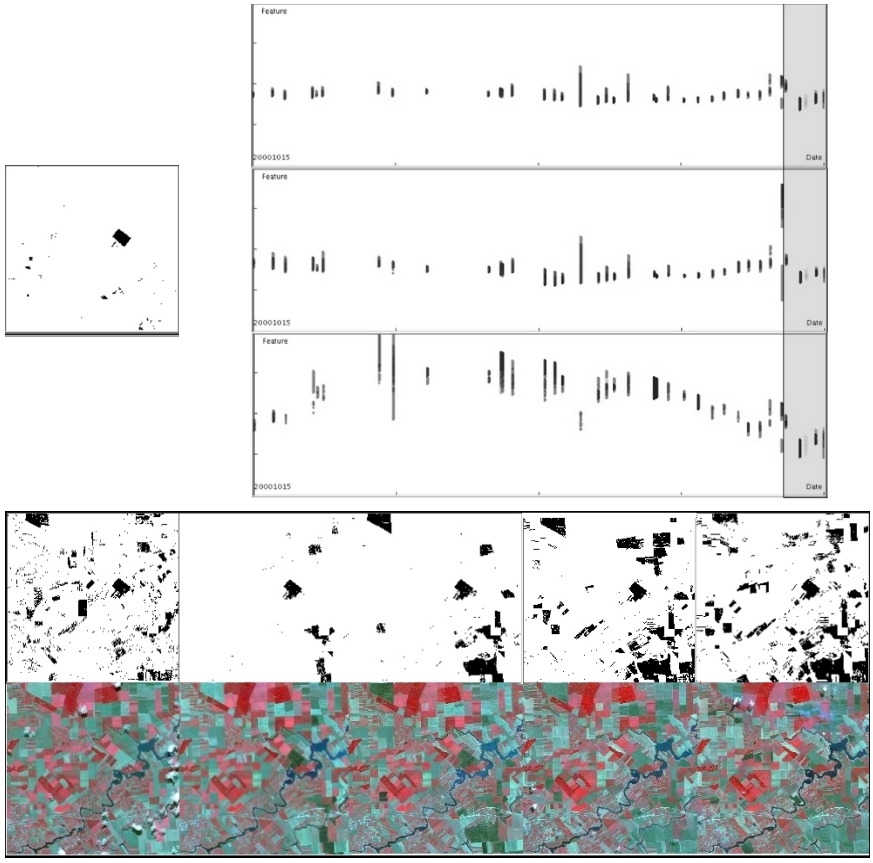


FIG. 8.7 – *Classe MT (en haut à gauche) ; projections des noeuds du graphe associés à la classe MT, sur les 3 différentes bandes spectrales (en haut à droite) ; classification dynamique associée à la classe MT et série temporelle d'images dans la fenêtre temporelle définie par la portion du graphe coloriée en gris (en bas).*

La qualité du processus d'inférence de graphe pourrait s'évaluer en comparant les trajectoires et classifications dynamiques obtenues avec des données réelles d'évolution recueillies sur le terrain. Mais de telles données ne sont pas accessibles et nous nous contentons de valider la procédure avec l'analyse visuelle d'un exemple de graphe. La figure 8.7 présente l'évolution d'un cluster associé à un champ de blé. On observe bien, entre le premier et le deuxième échantillon temporel contenu dans la zone grisée, que pour les 3 bandes spectrales, une rupture de la trajectoire du cluster, liée au labourage, est présente. Mais l'évaluation visuelle de la qualité des trajectoires est difficile. L'analyse de la classification dynamique associée est quant à elle plus intuitive. On se focalise sur la période de labours qui succède à la moisson, c'est à dire sur les 5 dernières images de la série. On observe que l'évolution spatiale des zones labourées est cohérente avec l'inspection visuelle des images. La première image de la classification dynamique présente la carte des champs de blé fraîchement moissonnés tandis que les images suivantes présentent pour chaque temps la carte des zones labourées.

Précisons que la qualité du graphe inféré dépend fortement des distributions des clusters engendrés dans l'étape précédente. En effet, des erreurs peuvent intervenir dans le calcul de l'intégrale de la divergence, lorsque l'on compare des gaussiennes de variances trop différentes[1]. Pour pallier ces erreurs de calculs, dans ce cas on approxime la variance la plus faible à un Dirac. En faisant l'hypothèse pour ce cas particulier d'indépendance entre les différentes bandes spectrales, l'intégration sujette à des erreurs de calcul s'approxime par une simple multiplication.

8.2.2 Compréhension de scènes dynamiques ADAM

Une évaluation exhaustive de la qualité de l'apprentissage sémantique reste à faire. Dans cette thèse, nous nous restreignons à valider par quelques exemples significatifs le bon fonctionnement et la puissance du sytème. On se focalise sur la compréhension de la dynamique spectrale de la scène. Afin de démontrer l'intérêt d'un tel système, capable d'exploiter la grande diversité d'information présente dans les séries temporelles, on présente dans la suite des exemples d'apprentissage de sémantiques variées, propres aux scènes dynamiques rurales.

Labourage de parcelles agricoles. On commence par la recherche de structures spatio-temporelles, associées au labourage de parcelles agricoles. On effectue l'apprentissage avec des exemples de structures spatio-temporelles définis dans une fenêtre spatiale de 200x200 pixels et dans une fenêtre temporelle constituée des 38 échantillons. Les structures recherchées sont définies spatialement par une classe MT et temporellement par une fenêtre constituée de 4 échantillons temporels. Notons qu'en choisissant la taille des fenêtres temporelles incluant les structures, on introduit déjà des connaissances a priori qui servent à écarter les structures candidates ne possédant pas cette échelle temporelle. Par apprentissage interactif de la sémantique, les probabilités a posteriori des structures sont inférées et, par maximisation de ces probabilités, les structures sont retrouvées dans l'espace et dans le temps. La figure 8.8 présente les structures retrouvées possédant les probabilités a posteriori les plus fortes. L'apprentissage a permis de retrouver la grande majorité des phénomènes de labours présents dans les données. On constate cependant quelques fausses détections et quelques détections de labours, possédant des probabilités relativement faibles. Notons que ces dernières ne sont pas présentées dans la figure 8.8.
On étend ensuite la recherche à une plus grande fenêtre spatiale de taille 800x800 pixels. La figure 8.9 présente les structures retrouvées possédant les probabilités a posteriori les plus fortes et les classifications dynamiques associées. On retrouve bien des structures de sémantique similaire. Notons que la visualisation de la classification dynamique permet de suivre l'évolution spatiale des labours, ce qui est particulièrement intéressant lorsque les tailles des images deviennent importantes.
Par ailleurs, remarquons que plus la fenêtre temporelle englobant les structures sera grande, plus les structures seront caractérisées et plus l'on pourra facilement différentier

[1] L'intégration numérique, employée pour le calcul analytique de la divergence entre gaussiennes, requiert une fonction à intégrer suffisamment lisse. Or, si les variances des gaussiennes ne sont pas du même ordre, la fonction divergence à intégrer ne sera pas lisse et des erreurs de calcul pourront apparaître.

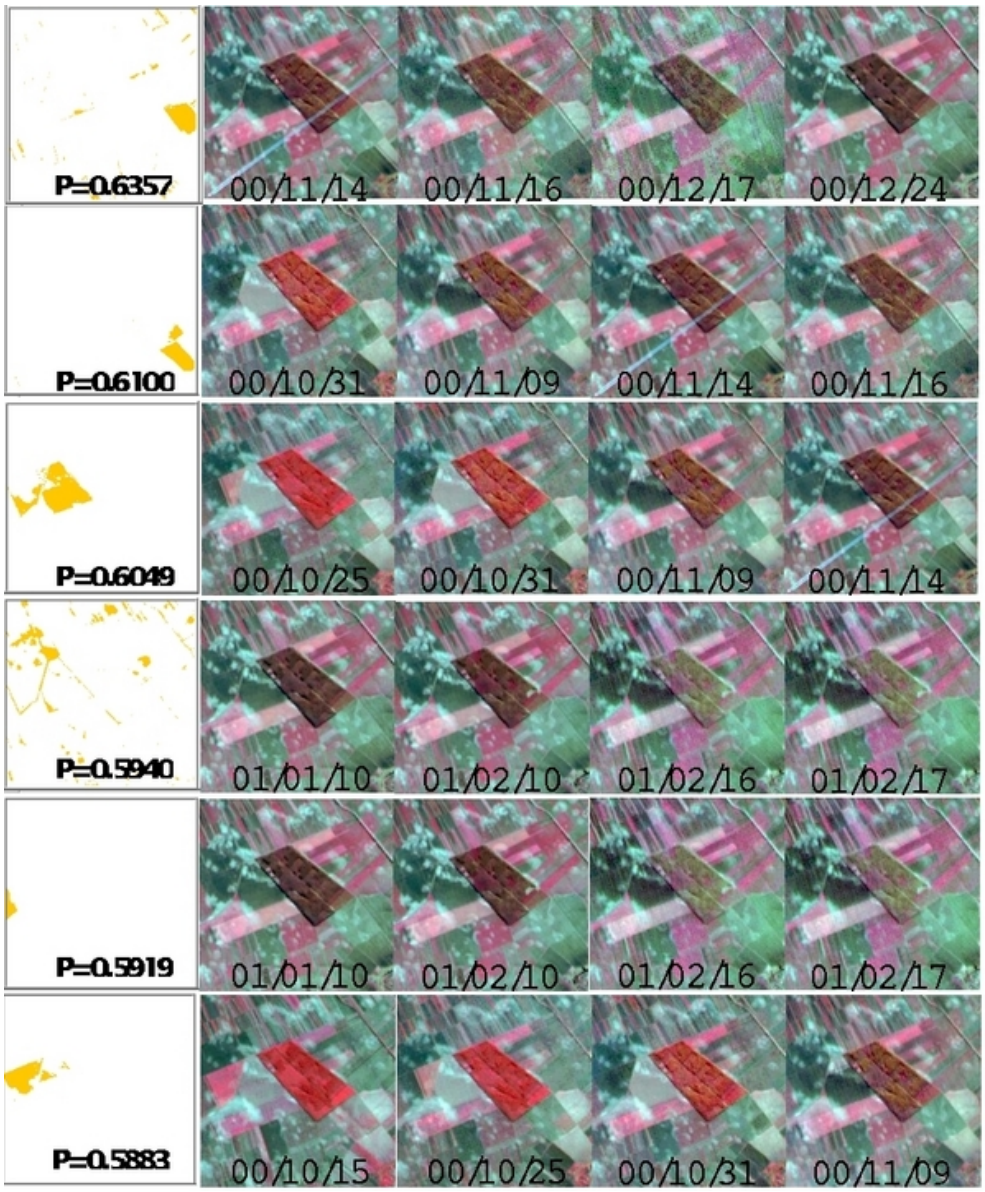


FIG. 8.8 – *Structures spatio-temporelles, associées au labourage de parcelles agricoles, retrouvées par apprentissage interactif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 200x200 pixels et dans une fenêtre temporelle constituée de 38 échantillons. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre constituée de 4 échantillons temporels, sont retrouvées dans l'espace et dans le temps.
La figure présente les structures spatio-temporelles les plus probables rangées, de haut en bas, en fonction de leurs probabilités a posteriori P. Chaque ligne présente une structure spatio-temporelle retrouvée, i.e. une classe MT retrouvée (gauche) avec sa fenêtre temporelle associée (donnée par les dates affichées dans la première et dernière image de la ligne), ainsi que la série temporelle d'images (droite) où la structure spatio-temporelle apparaît.

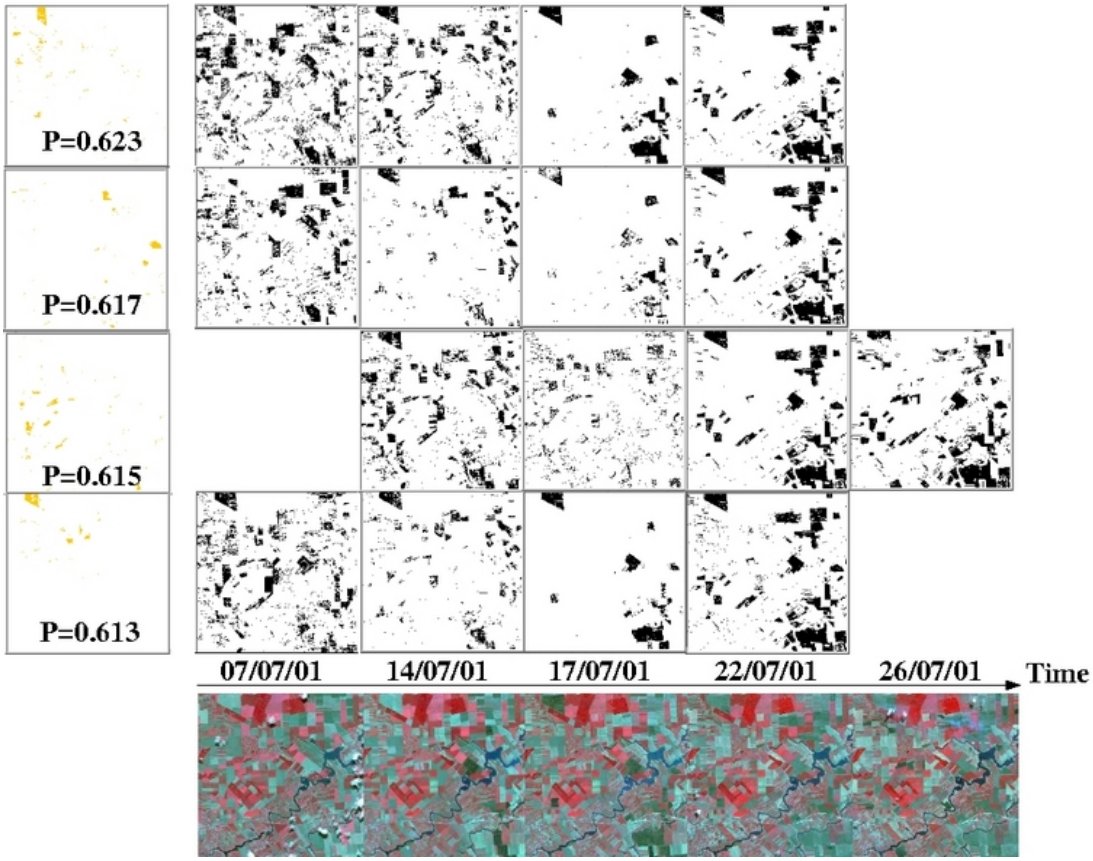


FIG. 8.9 – *Structures spatio-temporelles, associées au labourage de parcelles agricoles, retrouvées par apprentissage interactif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 800x800 pixels et dans une fenêtre temporelle constituée de 15 échantillons. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre constituée de 4 échantillons temporels, sont retrouvées dans l'espace et dans le temps.
La figure présente les structures spatio-temporelles les plus probables rangées, de haut en bas, en fonction de leurs probabilités a posteriori P. Chaque ligne présente une classe MT retrouvée (gauche) avec sa classification dynamique associée (droite). La fenêtre temporelle dans laquelle est définie la structure retrouvée est donnée par les localisations temporelles de la première et la dernière classification de la ligne. La série temporelle d'images, contenant l'ensemble des structures spatio-temporelles retrouvées, est présentée en bas de la figure.

leurs sémantiques. Mais inversement, une structure inclue dans une fenêtre temporelle trop importante poura être affectée par des perturbations importantes et pourra donc être confondue avec d'autres structures. En conséquence, une fenêtre temporelle ni trop grande, ni trop étroite, pourra engendrer une bonne discrimination sémantique. Les phénomènes de labours se déroulent sur une période relativement courte. Par ailleurs, ces phénomènes possèdent des caractéristiques variables en fonction de la nature du champ labouré. Par conséquent la différenciation d'une classe de sémantique positive d'une classe de sémantique négative, est délicate. On se limite donc à la définition de la sémantique positive. Il en découle des probabilités a posteriori peu élevées et une relativement faible discrimination de cette sémantique.
Cet exemple démontre les capacités du système à apprendre une sémantique très précise, à savoir le labourage après la moisson des champs de blé. Par ailleurs, plus que la seule recherche probabiliste de ces types de phénomènes, le système propose à l'utilisateur de comprendre l'évolution spatiale du processus de labourage via l'analyse visuelle des classifications dynamiques. Notons que, bien que les graphes de trajectoires ne soient pas présentés ici, l'utilisateur a aussi accès à l'analyse visuelle de l'évolution des caractéristiques au travers des représentations sous forme de trajectoires de clusters dynamiques.

Maturation de cultures parcellaires. On entraîne ensuite le système à retrouver des structures spatio-temporelles associées à la maturation de cultures parcellaires. Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 200x200 pixels et dans une fenêtre temporelle constituée de 38 échantillons. Comme les phénomènes de maturation se déroulent sur une longue période, on recherche des structures inclues dans une fenêtre temporelle de 12 échantillons. On communique quelques exemples positifs qui suffisent à la découverte de structures similaires. Contrairement à la sémantique de labours précédente, l'apprentissage d'une sémantique négative est effectué avec très peu d'exemples. Cette facilité d'apprentissage de la sémantique négative est vraisemblablement due à la largeur importante de la fenêtre temporelle considérée. En effet, les fenêtre temporelles sont d'une taille suffisante et permettent d'engendrer une bonne caractérisation et une bonne différentiation sémantique des structures. On obtient donc un bon apprentissage de la sémantique positive et de la sémantique négative. Il en découle que les structures retrouvées sont associées à de fortes probabilités a posteriori. Les résultats de la recherche dans l'espace et dans le temps sont présentés dans la figure 8.10. Les structures spatio-temporelles sont représentées avec des séries temporelles sous-échantillonnées (seulement trois images significatives ont été sélectionnées à partir des 12 échantillons temporels de la série). On note que les évolutions de champs associées à de fortes probabilités correspondent à la maturation de parcelles de soja, donnée pour exemple positif durant l'apprentissage, tandis que les évolutions associées à des probabilités moins fortes correspondent à des maturations de cultures différentes.
Cet exemple met en évidence l'aptitude du système à apprendre une sémantique très générale associée à des structures relativement différentes. En effet, la sémantique de maturation peut aussi bien être associée à la culture spécifique du soja, qu'à l'accroissement de la masse foliaire d'une forêt.

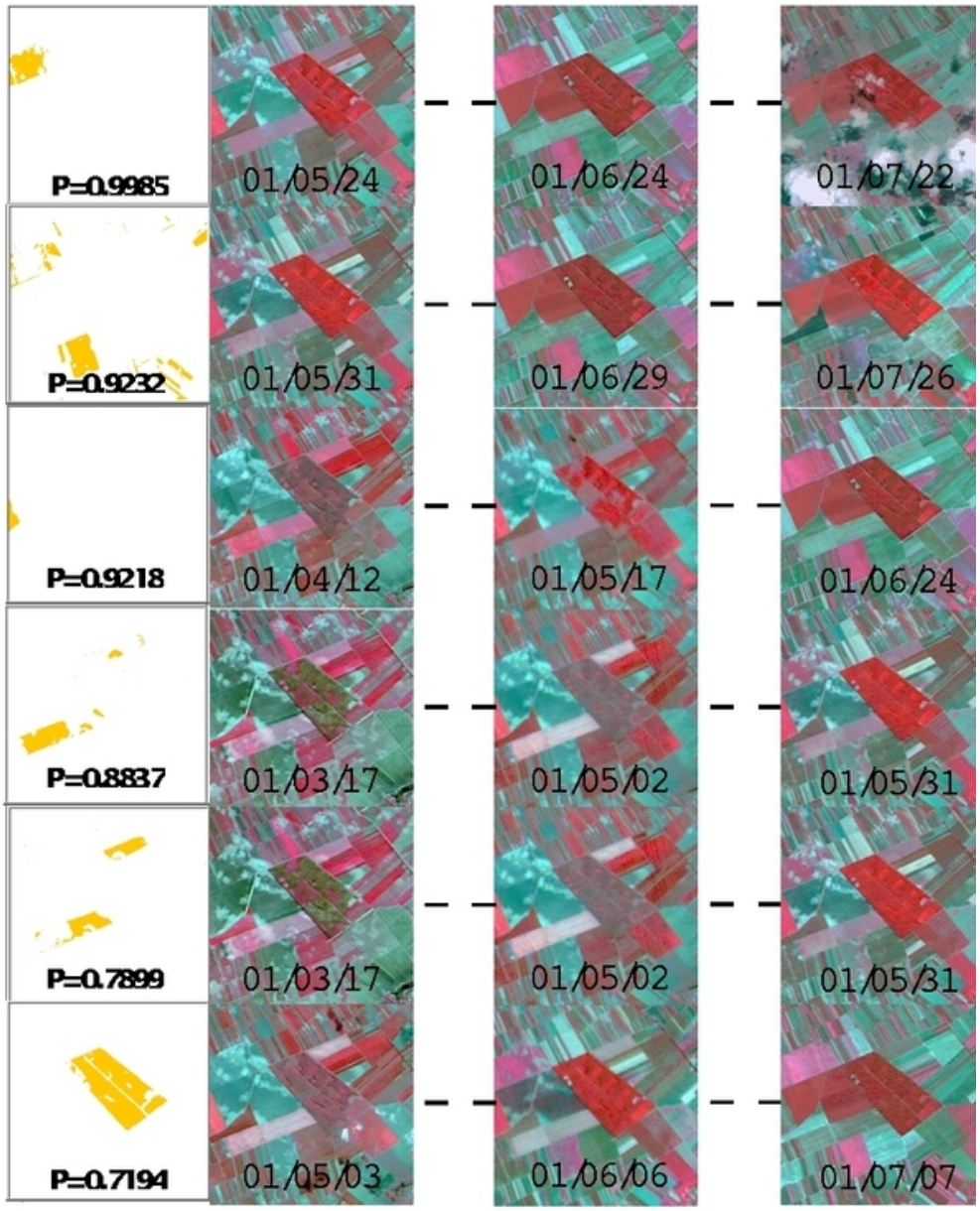


FIG. 8.10 – *Structures spatio-temporelles, associées à la maturation de cultures parcellaires, retrouvées par apprentissage interactif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 200x200 pixels et dans une fenêtre temporelle constituée de 38 échantillons. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre constituée de 12 échantillons temporels, sont retrouvées dans l'espace et dans le temps.
La figure présente les structures spatio-temporelles les plus probables rangées, de haut en bas, en fonction de leurs probabilités a posteriori P. Chaque ligne présente une structure spatio-temporelle retrouvée, i.e. une classe MT retrouvée (gauche) avec sa fenêtre temporelle associée (donnée par les dates affichées dans la première et dernière image de la ligne), ainsi que la série d'images (droite) temporellement sous-échantillonnées où la structure apparait.

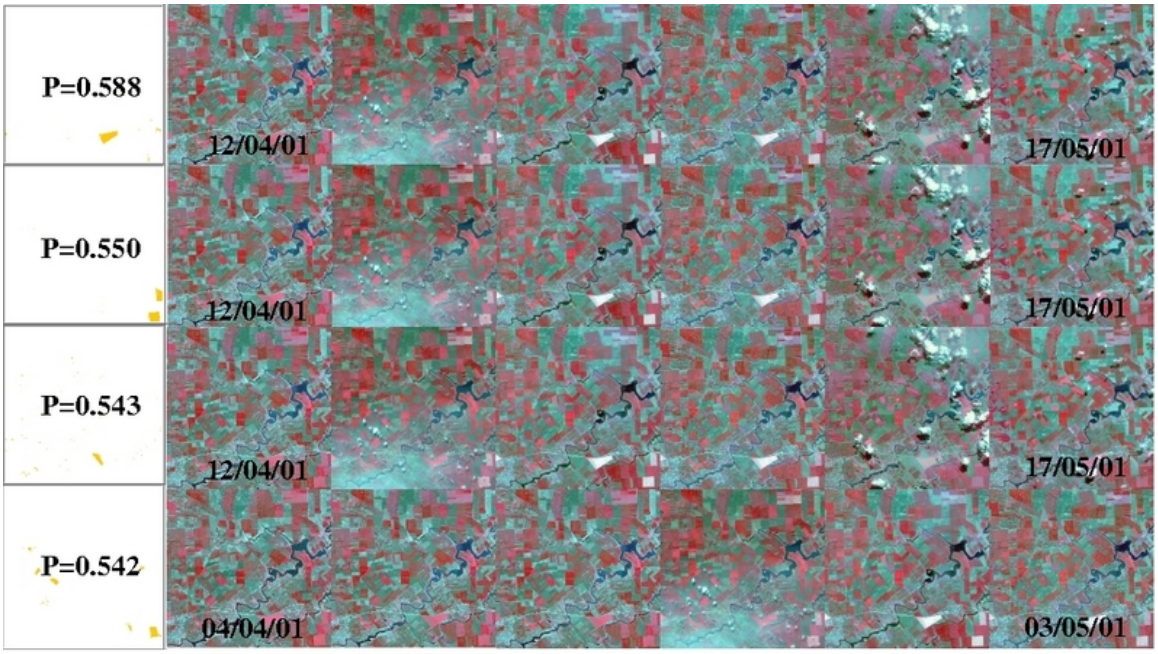


FIG. 8.11 – *Structures spatio-temporelles, associées à la floraison de parcelles de colza, retrouvées par apprentissage interactif avec un seul exemple positif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 800x800 pixels et dans une fenêtre temporelle comprenant les échantillons acquis entre mars et juin 2001. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre constituée de 6 échantillons temporels, sont retrouvées dans l'espace et dans le temps.
Les structures spatio-temporelles les plus probables sont rangées, de haut en bas, en fonction de leurs probabilités a posteriori P. Chaque ligne présente une classe MT retrouvée (à gauche) avec sa fenêtre temporelle associée (donnée par les dates affichées dans la première et dernière image de la ligne).

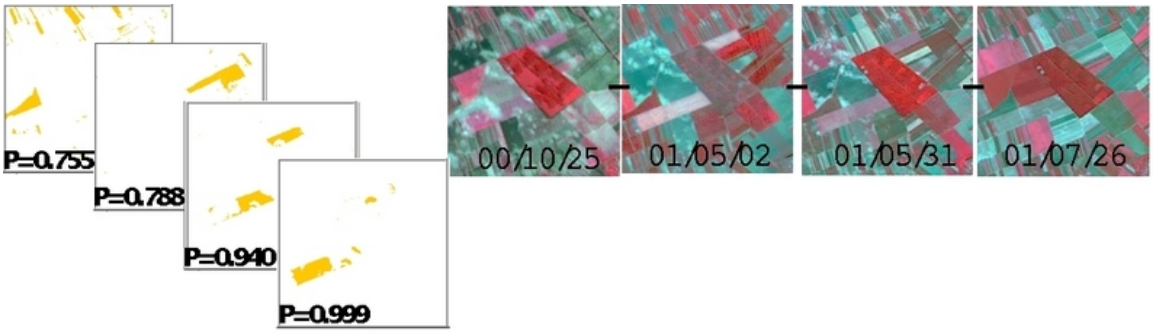


FIG. 8.12 – *Structures spatio-temporelles, associées à la floraison de parcelles de colza, retrouvées par apprentissage interactif avec un seul exemple positif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 200x200 pixels et dans une fenêtre temporelle comprenant les 38 échantillons. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre temporelle maximale constituée de 38 échantillons, sont retrouvées dans l'espace.
Les structures spatio-temporelles les plus probables sont rangées, du premier plan au dernier plan, en fonction de leurs probabilités a posteriori P. Chaque plan présente une classe MT retrouvée (à gauche). Les images affichées (à droite) représente la série temporellement sous-échantillonnées qui contient l'ensemble des structures spatio-temporelles.

Evolution et floraison du colza. On se focalise maintenant sur l'apprentissage d'une sémantique associée à la floraison du colza. La période de floraison du colza est relativement brève, quelques semaines, et se situe au printemps. En conséquence, les structures recherchées sont inclues dans des fenêtres temporelles constituées de 6 échantillons compris entre mars et juin. Un unique exemple communiqué au système suffit à retrouver les floraisons similaires dans une fenêtre spatiale de 800x800 pixels. La figure 8.11 présente les structures spatio-temporelles de plus fortes probabilités retrouvées après la communication d'un exemple positif au système. On remarque, qu'à défaut d'exemples négatifs, les probabilités demeurent relativement faibles. La présence d'un aérosol est interprétée comme un phénomène de floraison.
Si l'on communique des exemples négatifs au système alors, la classe sémantique se différentie nettement. La figure 8.12 présente les champs de colza, retrouvés et clairement différenciés des autres cultures, recherchés dans une fenêtre d'une taille de 200x200 pixels pour une fenêtre temporelle de 38 échantillons.
Ce résultat est conforme aux remarques précédentes et conforte notre intuition : pour des fenêtres temporelles suffisament importantes, il est facile de discriminer des structures en fonction d'une certaine sémantique.

Pratiques culturales du blé et du pois. Enfin, on recherche des structures spatio-temporelles associées à des parcelles soumises à certaines pratiques culturales. On s'intéresse donc à des évolutions annuelles identiques et plus particulièrement au cycle du blé : en automne, le champ est labouré puis ensemencé ; les cultures végètent durant l'hiver ; au printemps, les cultures grandissent jusqu'à maturation ; à la fin de l'été le blé est finalement

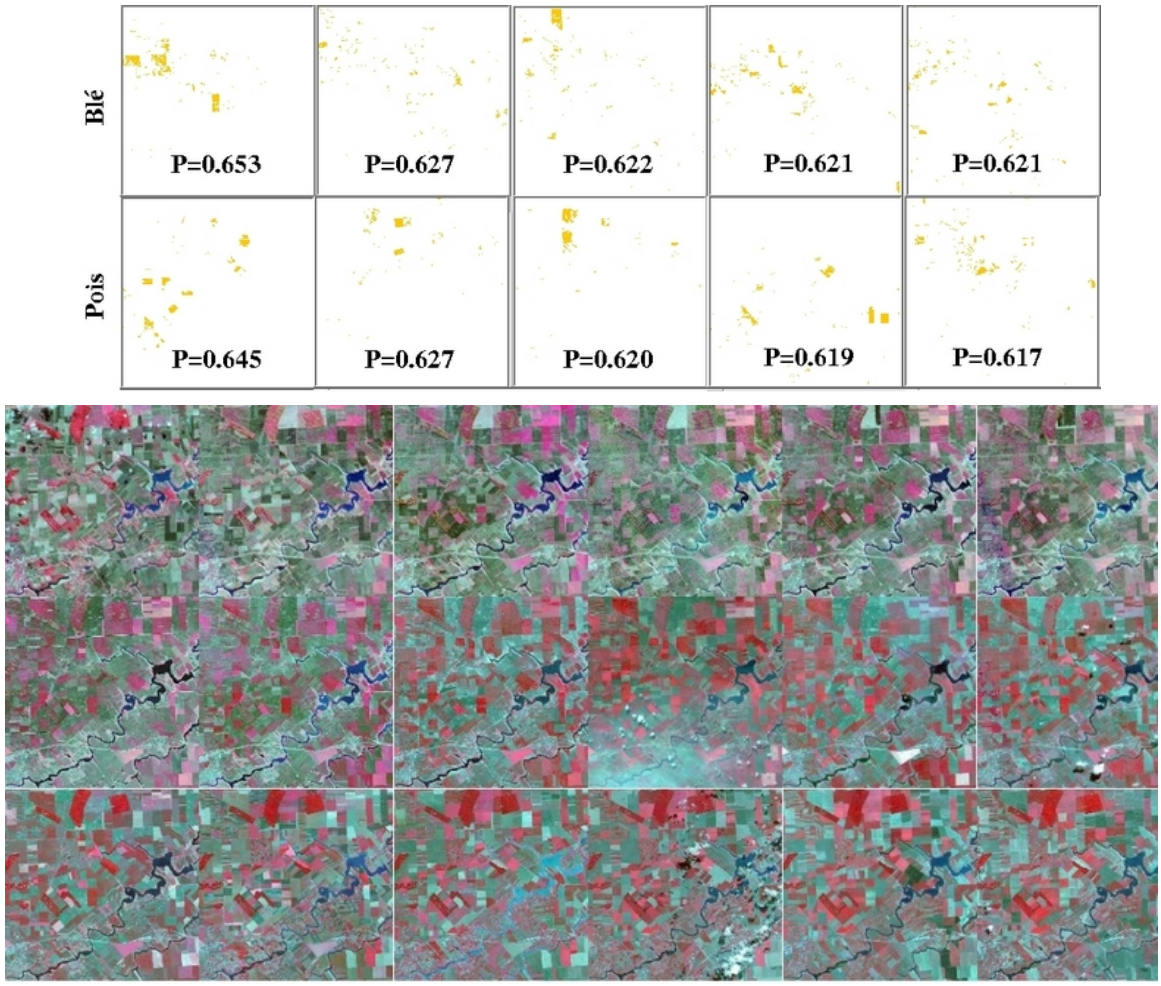


FIG. 8.13 – *Structures spatio-temporelles, associées à certaines pratiques culturales du blé ou du pois, retrouvées par apprentissage interactif avec un seul exemple positif.*
Les structures spatio-temporelles sont recherchées dans une fenêtre spatiale de 800x800 pixels et dans une fenêtre temporelle comprenant les 38 échantillons. Les structures les plus probables, définies spatialement par une classe MT et temporellement par une fenêtre temporelle maximale constituée de 38 échantillons, sont retrouvées dans l'espace. Un exemple d'évolution de parcelle de blé ou de pois, communiqué au système, permet la découverte de parcelles de même nature ayant subi des pratiques culturales similaires (même période de récolte, labours, etc).
La figure présente sur deux lignes, correspondant aux 2 différents apprentissages, les structures spatio-temporelles les plus probables (en haut). Dans chaque ligne, les classes MT retrouvées sont rangées, de gauche à droite, en fonction de leurs probabilités a posteriori P. La série temporelle d'images, sous échantillonnée temporellement, est présentée sur trois lignes (en bas). Celle-ci contient l'ensemble des structures spatio-temporelles retrouvées.

moissonné. On s'intéresse aussi à l'évolution des cultures de pois : l'évolution se caractérise par le développement des feuilles et des ramifications au printemps, une floraison début juin et une récolte en août. On recherche les structures spatio-temporelles dans une fenêtre spatiale de 800x800 pixels. Dans cette fenêtre spatiale, on choisit un seul exemple, défini spatialement par une classe MT et temporellement par une fenêtre temporelle maximale constituée de 38 échantillons. On retrouve ainsi les parcelles de nature similaire qui ont subi des changements identiques. Par exemple, on pourra différencier des parcelles de nature similaire, qui ont été moissonnées à des périodes différentes.
Les structures les plus probables retrouvées après avoir fourni au système un exemple particulier de champs de blé ou de champs de pois, sont présentées dans la figure 8.13. D'abord, on observe une répartition spatiale très éparse des structures associées à ces sémantiques. Ensuite, pour comprendre pourquoi ces structures ont été retrouvées avec les plus fortes probabilités a posteriori, il nous faut inspecter minutieusement l'évolution temporelle des structures représentées dans la série temporelle d'images. Cette inspection est menée et nous permet de conclure qu'en effet des pratiques culturales similaires caractérisent les parcelles retrouvées. En conséquence, cet exemple démontre la capacité du système à reconnaître des phénomènes complexes, dispersés dans l'espace et marqués dans le temps par des événements similaires. Pour arriver au même résultat avec la seule inspection visuelle, des ressources humaines considérables auraient dû être employées.

Au travers de ces exemples d'étiquetage sémantique, la puissance du système a été mise en évidence. La variété de l'information contenue dans les séries temporelles d'images satellites est en effet exploitée. La sémantique modélisée peut aussi bien être associée à un phénomène très spécifique qu'à des évolutions plus générales. Par ailleurs, ces exemples ont démontrés les performances du système pour rechercher des structures dans l'espace et dans le temps et les comprendre. Toutefois, on note des difficultés d'apprentissage d'une sémantique négative, i.e. de discrimination d'une sémantique positive, lorsque l'on emploie des fenêtres temporelles trop petites pour caractériser les structures.

8.3 Résumé

Dans ce chapitre, les points suivants ont été considérés.

- Le concept de fouille spatio-temporelle, présenté dans les chapitres précédents, a été intégré dans un système possédant une architecture client-serveur ; les composantes et les fonctionnalités de l'interface Homme-Machine ont été notamment détaillées. L'implémentation actuelle des algorithmes de la chaîne de traitement a été présentée et de possibles améliorations futures ont par ailleurs été précisées.

- Les algorithmes et techniques composant le système ont été évalués et validés avec plusieurs scènes dynamiques extraites de la base de données ADAM ; en particulier, des comparaisons avec des données "terrains" et des exemples d'étiquetage sémantique ont démontré la puissance de l'apprentissage non-supervisé et interactif.

- Ce système de fouille possède une approche exploratroire novatrice dans le domaine ; son adaptation aux intérêts des utilisateurs, permet l'exploitation efficace de l'information complexe contenue dans les séries temporelles d'images satellites.

Chapitre 9

Conclusion

Dans cette thèse, un concept a été proposé pour l'apprentissage de structures spatio-temporelles contenues dans les séries temporelles d'images satellites. Les méthodes développées ont été intégrées dans un système dédié à l'exploitation des séries temporelles d'images satellites, comprenant d'une part une chaîne de traitement automatique, et d'autre part un module pour l'étiquetage sémantique interactif des structures spatio-temporelles.

9.1 Résumé des travaux

Le concept développé se propose de résoudre le problème complexe de l'application du paradigme de la fouille d'information aux séries temporelles d'images satellites de haute résolution. L'approche de fouille d'information est motivée par les limitations des techniques classiques d'analyse des séries temporelles employant des modèles physiques face à l'accroissement de la quantité d'information disponible lorsque l'on atteint des résolutions de l'ordre du décamètre.

Les techniques développées s'inscrivent dans le contexte de l'analyse bayésienne. Ces méthodes d'analyses, ainsi que leurs équivalences avec des méthodes entropiques, sont largement détaillées.

Un modèle hiérarchique bayésien est développé pour lier les structures spatio-temporelles à des sémantiques d'utilisateurs. Le flux d'information entre un utilisateur et les données spatio-temporelles est modélisé hiérarchiquement. Ainsi, des niveaux croissants d'abstraction de l'information sont atteints par inférence bayésienne ou entropique.
La représentation hiérarchique de l'information engendrée par modélisation bayésienne est compatible avec le concept implémenté par le système de fouille d'information dans les images I2M (c.f. Annexe A.1). Dans la continuité de ce concept, cette thèse étend la modélisation hiérarchique des structures spatiales, à celle des structures spatio-temporelles contenues dans les scènes dynamiques.

Globalement, la modélisation hiérarchique se décompose en un apprentissage non-supervisé et un apprentissage interactif. Plus précisément, une modélisation bayésienne est adaptée à chaque niveau de la hiérarchie. Pour y parvenir, cette thèse propose les nouvelles méthodes suivantes :

- des caractéristiques primitives d'images sont extraites et analysées pour donner naissance à deux types de représentations complémentaires : la représentation multitemporelle (MT) et les représentations temporellement localisée (TL) ;

- le principe MDL est appliqué pour inférer un modèle de mélange d'un nombre indéterminé de gaussiennes corrélées ; ce modèle est utilisé pour apprendre les différentes distributions de caractéristiques ;

- la complémentarité des représentations TL et MT est exploitée pour l'inférence de trajectoires de clusters dynamiques ; l'information complexe des structures spatio-temporelles est ainsi condensée dans les noeuds et les arcs d'un graphe de trajectoire ;

- basé sur le graphe, un apprentissage interactif bayésien est développé pour effectuer l'étiquetage sémantique des structures spatio-temporelle.

Détaillons plus amplement ces différentes méthodes.

Les méthodes d'extraction de caractéristiques primitives d'images sont directement importées du concept I2M. Elles engendrent d'une part une représentation MT. Cette dernière contient l'information spatio-temporelle globale sur les structures. Mais la modélisation de l'espace MT des caractéristiques étant difficile du fait de sa dimension élevée, des techniques linéaires et non-linéaires d'analyse multidimensionnelle sont appliquées pour réduire la dimension de cet espace. Toutefois, parce que ces techniques possèdent des limites d'analyse, on utilise d'autre part les représentations TL pour une extraction d'information plus détaillée.

L'hypothèse du modèle de mélange d'un nombre indéterminé de gaussiennes corrélées est appliquée pour apprendre les distributions des caractéristiques dans les espaces MT et TL. L'inférence des paramètres et de la complexité du modèle se base sur le principe MDL. Les équivalences de ce principe avec les méthodes bayésiennes ont par ailleurs été démontrées. Pour parvenir au minimum de la longueur de description, et ainsi sélectionner un modèle et estimer les paramètres de celui-ci, une méthode de plus forte descente est utilisée en association avec un algorithme EM. Pour chaque représentation, la modélisation par mélange de gaussiennes permet le regroupement des caractéristiques et des pixels respectivement en clusters et classes.

La distribution de caractéristiques évolutives est modélisée sont forme de trajectoire de clusters dynamiques. Pour obtenir ces trajectoires, les clusters MT, assimilables à des

clusters dynamiques, sont d'abord projetés aux différents instants. Puis, leurs similarités avec les clusters TL sont évaluées. Notons que la similarité est évaluée conjointement dans l'espace de caractéristiques et dans l'espace euclidien par l'utilisation de la distance de Kullback-Leibler entre clusters pondérée par une surface de recouvrement de classes. Les mesures de similarités servent à engendrer des probabilités d'associations entre cluster MT et TL. On formalise ainsi les trajectoires des clusters MT au travers des clusters TL.
Un graphe, regroupant des trajectoires de clusters MT est alors formé. Dans ce graphe, les trajectoires des clusters MT sont associées à des noeuds représentant les clusters TL et à des arcs représentant l'évolution des clusters TL entre échantillons temporels consécutifs. Des attributs tels que les paramètres des clusters, le flux de points ou encore l'information mutuelle entre deux états consécutifs, sont associés aux noeuds et aux arcs du graphe de trajectoires De plus, les classes TL associées aux différents noeuds permettent la génération de classifications dynamiques. Celles-ci sont représentatives de l'évolution spatiale des structures. Ce graphe constitue le résultat de la procédure d'apprentissage non-supervisé.

L'apprentissage interactif de sémantiques s'appuie sur la représentation des structures spatio-temporelles sous formes de graphes de trajectoires. Une inférence a posteriori est utilisée pour dériver, par l'intermédiaire d'exemples positifs et négatifs fournis par des utilisateurs, les probabilités d'associations des structures avec les différents labels sémantiques. Le modèle sémantique se base sur un modèle paramétrique de similarité dérivé de l'algorithme d'appariement inexact. L'estimation des paramètres du modèle est effectuée via l'apprentissage interactif des hyper-paramètres d'un modèle de Dirichlet appliqué à un réseau bayésien. Par ailleurs, des mesures de pertinence, mises à jour à chaque itération du processus d'apprentissage, permettent l'évaluation de la qualité des labels sémantiques définis.

Les méthodes développées sont intégrées dans un système de fouille visuelle d'information à l'architecture client-serveur, qui emploie une interface Homme-Machine pour superviser l'apprentissage sémantique. Plusieurs représentations de l'information sont visualisées au travers de cette interface graphique. En particulier, les utilisateurs peuvent naviguer dans le graphe de trajectoires ainsi qu'au travers des classifications dynamiques générées.

Les algorithmes et techniques composants le système ont été évalués avec des scènes dynamiques extraites de la base de données ADAM. En particulier, des comparaisons avec des données "terrains" révèlent la pertinence de l'apprentissage non-supervisé et des exemples d'étiquetage sémantique valident l'apprentissage supervisé.

Le système développé dans cette thèse constitue une première approche prometteuse pour étendre le paradigme de la fouille d'information aux séries temporelles d'images satellites. L'approche exploratoire et spécifique aux intérêts des utilisateurs est novatrice dans le domaine. Contrairement aux analyses de séries temporelles classiques qui se restreignent à la compréhension de certaines structures, cette approche permet l'exploitation efficace de la grande diversité de l'information contenue dans de telles données.

9.2 Perspectives

Plusieurs axes intéressants de recherche et développement se dégagent de ces travaux de thèse.

D'abord, certains algorithmes du système actuel pourraient être codés de façon optimisée pour s'affranchir de certaines limitations. Par exemple, l'implémentation des algorithmes d'apprentissage sémantique et de l'interface Homme-Machine sous forme de servlets et d'applets est une condition nécessaire pour un fonctionnement en ligne du prototype. Par ailleurs, pour le processus de clustering, l'implémentation actuelle se place dans un cas particulier de modèle de mélange de gaussiennes et il est nécessaire d'étendre l'implémentation au cas général pour aller au delà de ces hypothèses limitatrices.

Ensuite, l'évaluation des performances d'un système de fouille n'est pas triviale. La validation des résultats obtenus par l'apprentissage de structures spatio-temporelles contenues dans les données ADAM ne constitue qu'une ébauche d'une procédure d'évaluation intensive, qui devra être effectuée avec de plus grandes quantités de données de nature variée et avec des caractéristiques diverses. Dans la perspective adoptée par *Datcu et Seidel (2002)*, cette procédure d'évaluation pourra s'appuyer sur des mesures entropiques puissantes telles que l'information mutuelle ou la complexité de Kolmogorov qui serviront à évaluer la perte d'information à chaque nouveau niveau d'abstraction décrit par le modèle hiérarchique. La robustesse du système face à la variabilité des données devra aussi être évaluée.

Enfin, des recherches théoriques pourront contribuer à l'amélioration du concept. Le développement d'un critère pour une sélection moins arbitraire du nombre de composantes indépendantes qu'engendre une analyse par PP de l'espace MT des caractéristiques, contribuerait à l'amélioration de la classification MT, sur laquelle toute l'analyse est centrée. Ou simplement, pour éviter l'analyse complexe de l'espace MT des caractéristiques, il serait intéressant d'imaginer de nouvelles méthodes pour l'inférence des trajectoires de clusters dynamiques, qui modéliseraient explicitement les processus temporels induits par l'évolution des clusters TL.
Par ailleurs, l'exploitation de l'information sur l'évolution géométrique est cruciale pour parvenir à une compréhension complète des phénomènes. Cette information géométrique peut bien sûr être extraite au premier niveau du modèle hiérarchique, c'est à dire durant l'étape d'extraction de caractéristiques. Cependant, cette extraction d'information est plus aisée au niveau des classifications dynamiques et l'extraction de caractéristiques de surface, de connexité, etc, pourraient s'ajouter facilement aux attributs actuels des graphes de trajectoires. Ainsi, par l'intermédiaire du modèle paramétrique de similarité, l'information géométrique pourrait être fusionnée aux autres informations. Aussi, une solution pour ne plus traiter indépendamment les caractéristiques spectrales et de texture, serait d'ajouter, relativement à chaque nature de caractéristique, des termes au modèle paramétrique de similarité On pourrait ainsi fusionner les graphes engendrés avec des caractéristiques de différentes natures lors de l'apprentissage sémantique.

Concernant l'apprentissage supervisé, le modèle sémantique pourrait être amélioré, afin de mieux s'adapter à la recherche de structures inclues dans des petites fenêtres temporelles, et ainsi aboutir à une meilleure discrimination sémantique.
Finalement, il serait intéressant d'étendre la recherche probabiliste de structures spatio-temporelles de sémantique similaire, à la recherche probabiliste de séries temporelles entières par leurs contenus sémantiques.

Annexe A

Le système de fouille d'information I2M et l'apprentissage de structures spatio-temporelles

A.1 Description du système

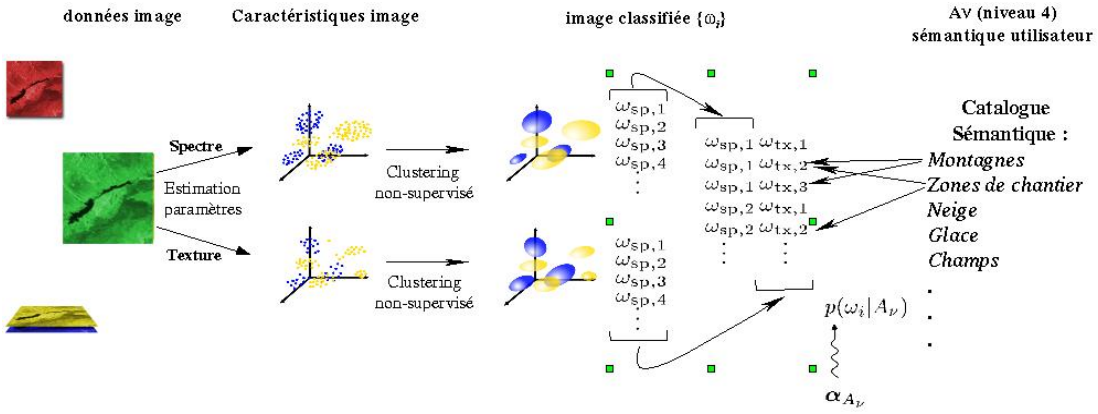


FIG. A.1 – *Modélisation hiérarchique de la description stochastique du contenu de l'image du système I2M.*

Le système de fouille d'information d'images *I2M* est conduit par les connaissances d'un utilisateur qui communique son intérêt pour des structures de l'image spatialement localisées au niveau de groupe de pixels. Cette perspective permet aussi bien la recherche par le contenu d'images entières, que la recherche de structures précises de l'image ou la compréhension détaillée de la scène.

Afin de construire un système qui ne dépende pas de la spécificité de l'application, le système *I2M* se base sur une modélisation hiérarchique du contenu des images *(Datcu et al., 1998 ; Schroder et al., 2000)*. La figure A.1 schématise ce

modèle. Le concept de représentation de l'information par des niveaux hiérarchisés d'abstraction sémantique se place dans le cadre de l'apprentissage bayésien. On distingue 4 niveaux de représentation.

- Premièrement, des caractéristiques primitives spectrales et texturales **(niveau 2)** sont extraites des données images **(niveau 1)** par l'utilisation de différents modèles paramétriques de signaux. En particulier, les champs aléatoires de Gibbs-Markov sont employés pour extraire des caractéristiques de texture à différentes résolutions. Des méta-caractéristiques, telles que l'évidence bayésienne du modèle employé pour un certain jeu de données considéré, sont ajoutées à la collection de caractéristiques primitives ;

- Ensuite, par un clustering non-supervisé des caractéristiques extraites, des cartes de classification des images sont obtenues pour chaque type de modèle d'extraction de caractéristiques employé, i.e. un vocabulaire de classes de signaux **(niveau 3)** est généré spécifiquement à chaque type de modèle d'extraction de caractéristiques. A partir des cartes de classification, les probabilités d'occurrence des différentes classes de signaux sont calculées pour chaque image. Les classifications et les histogrammes d'occurrence sont stockés dans une base de donnée relationnelle.

- Finalement, les intérêts spécifiques des utilisateurs, i.e. des labels sémantiques de type de couverture **(niveau 4)**, sont liés à des combinaisons de ces différents vocabulaires par l'utilisation d'un réseau bayésien. La modélisation sémantique présente des similarités avec celle développée dans ces travaux de thèse. Plus précisément, par la communication au système d'exemples positifs et négatifs de type de couverture, l'interêt subjectif d'un utilisateur pour certaines classes de signaux est mis en correspondance avec certaines probabilités du réseau bayésien. Ces probabilités sont inférées par l'intermédiaire d'un modèle de Dirichlet. Par ailleurs, une mesure de pertinence basée sur la divergence de Kullback-Leibler, guide l'apprentissage. Elle représente comment une sémantique dite positive, c'est à dire définie avec les exemples positifs, est séparée d'une sémantique dite négative, c'est à dire définie avec les contre-exemples. Cette mesure permet l'évaluation de la qualité des labels sémantiques.

Les labels sémantiques sont engendrés par seuillage des probabilités a posteriori associées aux différentes structures. Cet étiquetage des structures permet de calculer, pour chaque image, la probabilité a posteriori qu'elle contienne un certain label sémantique. Ainsi, par maximisation de ces probabilités a posteriori d'un critère de séparabilité ou de surface de recouvrement, une recherche probabiliste d'images est menée.

L'extraction de caractéristiques et le clustering sont des procédures, obtenues de manière complètement non-supervisée lors de l'ingestion des données dans le système, et ne dépendent pas de la spécificité de l'application. La représentation de l'information dans le niveau le plus haut du modèle hiérarchique est interactivement définie par les utilisateurs dans le cadre d'un paradigme d'apprentissage liant les classes de signaux (représentation objective) aux labels sémantiques d'utilisateurs (représentation subjective).

Ce système a démontré son efficacité dans l'extraction d'information dans plusieurs bases de données multi-capteurs et multirésolutions. De par la modélisation objective du contenu de l'image et la modélisation subjective de la sémantique d'un utilisateur, le système permet non seulement la fouille de larges bases de données, mais aussi la pleine compréhension des images. Un démonstrateur du système est accessible en ligne [94].

A.2 Adaptation du système aux données multitemporelles

La problématique est d'adapter le système à l'apprentissage de structures spatio-temporelles contenues dans une série temporelle d'images satellites de haute résolution. Une fenêtre spatiale de la série temporelle ADAM est présentée dans la figure A.2.

L'application d'une fenêtre temporelle glissante sur les séries temporelles d'images permet la génération de classifications de caractéristiques d'images temporellement localisées. En considérant que les fenêtres temporelles glissantes délimitent différentes séries temporelles d'images, le concept I2M peut s'adapter à la recherche de structures évoluant similairement dans celles-ci. L'intégration de données multitemporelles dans le système peut donc être assez directe, la seule adaptation étant l'élaboration de classifications multitemporelles issues des analyses successives de paquets de trames de la série *(Héas et al., 2003b)*.

Les classifications multitemporelles sont engendrées par les premiers niveaux du modèle hiérarchique bayésien décrit dans cette thèse. L'extraction de caractéristiques d'images, suivie par une procédure de réduction de dimensionnalité, précèdent donc la phase de classification (c.f. 6).

Une portion spatiale de la série temporelle ADAM, de taille 500x500 pixels, est utilisée pour expérimenter un premier cas relativement simple, où la fenêtre temporelle glissante ne contient qu'une seule image. Les caractéristiques spectrales et de texture extraites des 38 images de la série temporelle sont donc analysées par deux processus indépendant de classifications. Après l'ingestion dans la base de données du système des 38x2 cartes de classifications et des histogrammes d'occurrence associés, des labels sémantiques d'évènements apparaissant sur les images de la série peuvent être interactivement définis. La figure A.3 présente le résultat de l'apprentissage et de la recherche probabiliste d'une sémantique d'apparition de nuages.

La même série temporelle ADAM est utilisée pour expérimenter le cas, où la fenêtre temporelle glissante possède une taille maximum, c'est à dire contient 38 images. Des caractéristiques spectrales et de texture sont donc dans un premier temps extraites des 38 images de la série temporelle. Ces caractéristiques sont ensuite placées dans les espaces multitemporelles des caractéristiques. Ces espaces de grande dimensionnalité sont alors réduits par des techniques d'analyse multidimensionnelle. Une classification multitem-

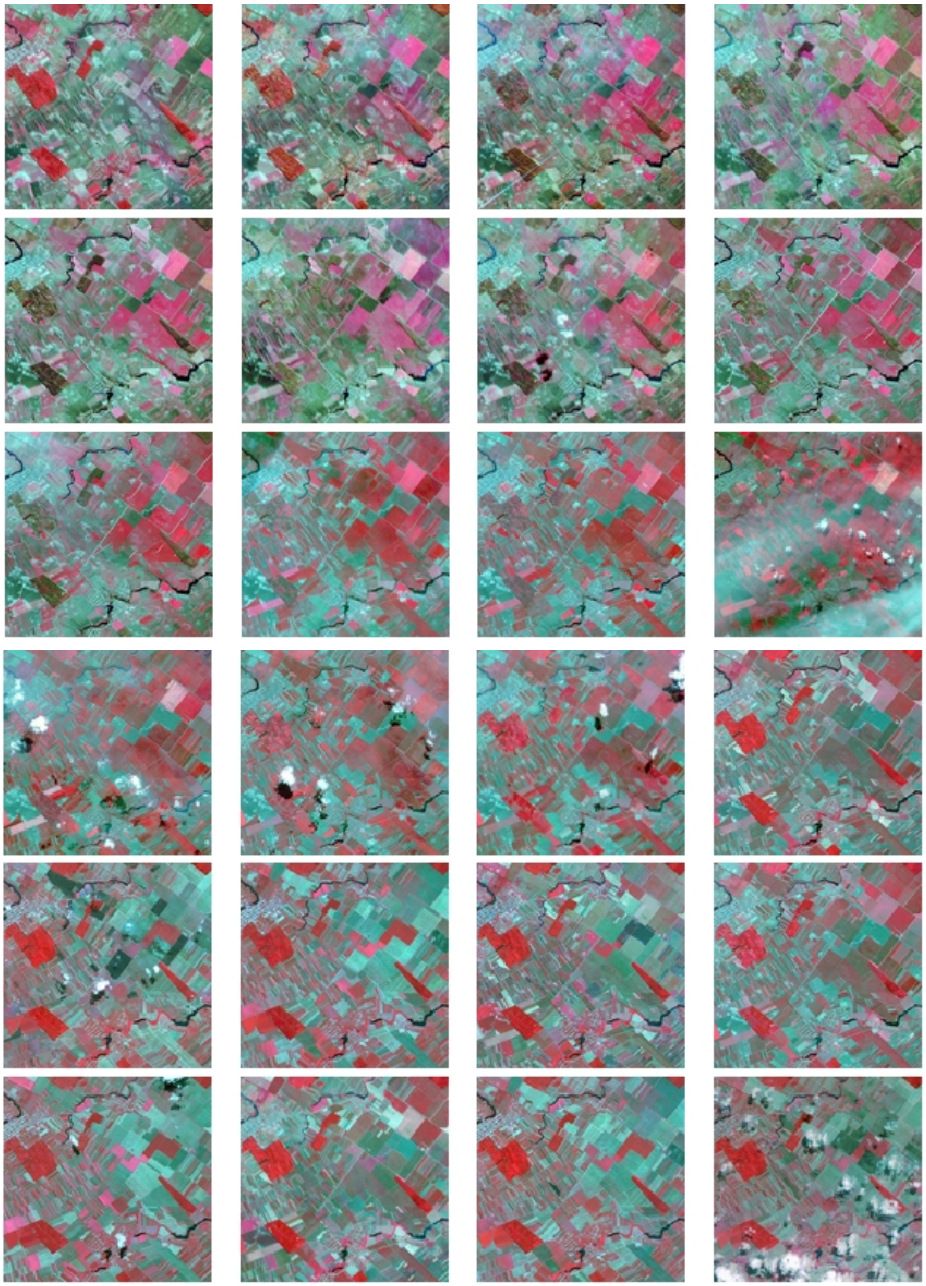


FIG. A.2 – *Fenêtre spatiale de 500x500 pixels de la série temporelle ADAM. Sur les 38 images de la série 24 images ont été sélectionnées pour représenter l'évolution de la série temporelle entière.*

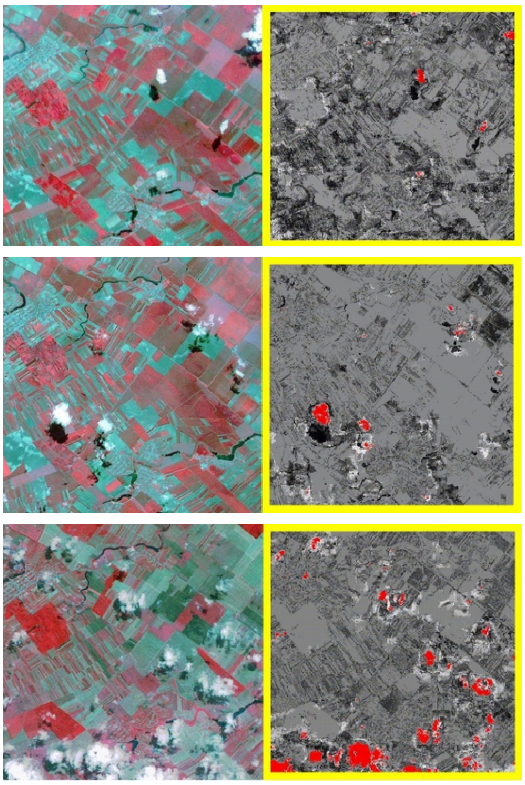


FIG. A.3 – *Etiquetage sémantique associé à une sémantique d'occlusion par des nuages.*

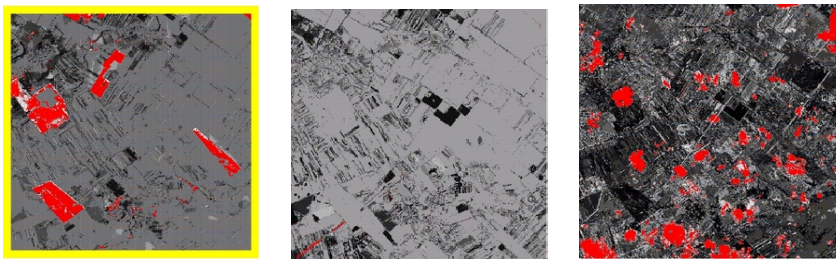


FIG. A.4 – *étiquetages associés à une sémantique d'évolution forestière (à gauche), d'occlusion à un temps donné par la traînée d'un avion (milieu) et par des nuages (à droite).*

porelle spectrale ainsi qu'une classification multitemporelle texturale sont engendrées et ingérées dans la base de données avec leurs histogrammes d'occurrence associés.
Par apprentissage interactif, trois différents étiquetages sémantiques d'évolution de couverture sont crées. La figure A.4 les présente. Un label pour des évolutions forestières similaires est d'abord appris par le système. Les classes de signaux spectraux et texturaux sont employées dans ce premier exemple. Notons que la zone forestière occultée par un nuage n'est pas associée au même label sémantique. Les deux autres exemples présentent des labels associés à des sémantiques d'occlusion. Ces exemples n'utilisent que des classes spectrales pour l'apprentissage. Dans l'un des exemples, la sémantique de l'occlusion par la traînée d'un avion est apprise, tandis que dans l'autre exemple, toutes les zones occultées par des nuages en des instants temporels différents sont labellisés.

Cet apprentissage de structures spatio-temporelles s'avère performant. Cependant, la principale limitation de cette approche, est qu'elle peut difficilement considérer la grande diversité d'intervalles temporels dans lesquels les structures spatio-temporelles sont définies. En effet, pour prendre en compte toutes les échelles temporelles des phénomènes, il est nécessaire d'ingérer dans le système autant de classifications multitemporelles qu'il existe de possibilités de sélectionner une portion de la série temporelle avec une fenêtre temporelle de taille variable. D'où l'intérêt de l'approche développée dans le corps de cette thèse.

Bibliographie

[1] R. Agrawal, *Data Mining : A Performance Perspective*, IEEE trans. on Knowledge and Data Engineering, vol. 5, no. 6, pp. 914-925, 1993.

[2] H. Akaike, *Information Theory and the Extension of the Maximum Likelihood Principle*, in B.N. Petrov, and F. Csaki, (Eds.) Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973.

[3] A. Antoniadis, *Estimation : concepts statistiques de bases*, Ecole de printemps “De la séparation de sources à l’analyse en composantes indépendantes, méthodes, algorithmes et applications, Villard-de-Lans, France, 2001.

[4] C. M. Antunes, *Temporal Data Mining : an overview*, IST, Lisbon Technical University, Workshop on Temporal Data Mining, KDD2001, 2001.

[5] C.M. Bachmann, E.E. Clothiaux, J.W. Moore, D.Q. Luong, *Dynamically reconfigurable Projection Pursuit ensembles for cloud detection in AVIRIS imagery* IEEE International geoscience and remote sensing symposium (IGARSS95), pp 256-259, vol.1, 1995

[6] R.A. Baxter et J. J. Olivier, *LMDL and MML : Similarities and Differences*, Tech. Report 207, departement of Computer Science, Monash University, Clayton, Vic. 3168, Australia, 1994.

[7] K. Barnard et D. Forsyth, *Learning the Semantics of Words and Pictures*, International Conference on Computer Vision, vol 2, pp. 408-415, 2001.

[8] *Base de données ADAM*,
http ://medias.obs-mip.fr/adam/web/prog/serveur.html

[9] R.A. Becker, S.G. Eick, G.J. Wills, *Visualizing Network Data*, IEEE trans. on Visualization and Graphics, Vol. 1, No. 1, pp 16-28, 1995

[10] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, *The R*-Tree : An Efficient and Robust Access Method for Points and Rectangles*, Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 322-331, 1990.

[11] J.P. Benzecri et al., *L’analyse des données (tomes 1 et 2)*, Paris Dunod, 1973.

[12] D. Berndt, J. Clifford, *Finding Patterns in Time-Series : a Dynamic Programming Approach*, in U. Fayyad, G. Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, pp. 229-248, AAAI Press, 1996.

[13] J. Besag, *On the Statistical Analysis of Dirty Pictures (with discussion)*, Journal of the Royal Statistical Society, B, 48, 259-302, 1986

[14] H. Bischof, A. Leonardis et A. Selb, *MDL Principle for Robust Vector Quantization*, Pattern Analysis and Applications, 2 :59-72, 1999.

[15] J.S. Borecsky, L.A. Rowe, *Comparison of video shot boundary detection techniques*, Proc. of SPIE, vol. 26670, pages 170-179, 1996.

[16] L. Bruzzone and D.F. Prieto, *An Adaptative Parcel-Based Technique for Unsupervised Change Detection*, International Journal of Remote Sensing, 21(4ř :817-822, 2000a.

[17] L. Bruzzone and D.F. Prieto, *Automatic Analysis of the Difference Image for Unsupervised Change Detection*, IEEE Transactions on Geoscience and Remote Sensing, 38(3) : 1171-1182, 2000b.

[18] H. Bunke and G. Allerman, *Inexact graph matching for structural pattern recognition*, Pattern Recognition Lett. 1(4), pp. 245-253, 1983.

[19] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Tay- lor et D. Freeman, *AutoClass : A Bayesian classification system*, Proceedings of the Fifth International Conference on Machine Learning, Ann Arbor, MI., June 12-14, 1988.

[20] M.Ciucu, P. Heas, M. Datcu et J. C. Tilton, *Scale Space Exploration For Mining Image Information Content*, Revised Papers from the Third International Workshop on Multimedia Data Mining / Knowledge Discovery and Data Mining, Lecture Notes in Computer Science 2797, Springer 2003, ISBN 3-540-20305-2, pp 30-38, 2003.

[21] B.S. Clarke et A.R. Barron, *Information-theoric asymptotics of Bayes methods*, IEEE Trans. Inform. Theory 36 : 1197-1208, 1990.

[22] W.Cleveland, *Visualizing Data*, Summit, MJ. Hobart Press, 1993.

[23] M.P. Consens, A.O. Mendelzon, *Hy+ : A Hygraph-based Query and Visualization System*, Proc.ACM SIGMOD int. Conf. on Management of Data, Washington DC, pp 511-516, 1993.

[24] I. Cox, M. Miller, T. Minka, T. Papathomas et P. N. Yianilos, *The Bayesian Image Retrieval System, PicHunter : Theory, Implementation and Psychophysical Experiments* , IEEE Transactions on Image Processing, Vol 9, No 1, pages 20-37, Jan 2000.

[25] G.F. Cooper et E. Herskovits , *A Bayesian method for the induction of probabilistic networks from data*, Machine Learning, 9, 309-347

[26] T.M. Cover et J.A. Thomas, *Elements of Information Theory*, new York : John Wiley & sons, 1991.

[27] M. Datcu, K. Seidel et M.Walessa, *Spatial Information Retrieval from Remote Sensing Images - Part 1 : Information theoritical perspective*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 36, no. 5, pp. 1431-1445, 1998.

[28] M. Datcu, K. Seidel, *Bayesian methods : applications in information aggregation and data mining*, International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part 7-4-3 W6, pp. 68-73, 1999a.

[29] M. Datcu, *Information theoretical aspects and semantic representations in advanced communication systems*, Habilitation, University Louis Pasteur, Strasbourg, France, 1999b.

[30] M. Datcu, and K. Seidel, *KIM - Knowledge Driven Information Mining in Remote Sensing Image Archives*, http ://earth.esa.int/rtd/IIMCG/documents.html, November 12, 2002.

[31] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, S. D'Elia, *Information mining in remote sensing image archives : system concepts*, IEEE Transactions on Geoscience and Remote Sensing, Volume : 41, Issue : 12, pp 2923- 2936, 2003.

[32] S. Deerwester et al., *Indexing by Latent Semantic Analysis*, Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[33] A.P. Dempster, N.M. Laird et D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm* Journal of the Royal Statistical Society, B 39, pp 1-38, 1977.

[34] R. Duda, P. Hall, D.Stork, *Pattern Classification (2nd edition)*, Wiley-Interscience, 2000.

[35] J. P. Eakins, *Automatic Image Content Retrieval - Are We Getting Anywhere*, Proc. of the Third International Conf. on Electronic Librairy and Visual Information Research, De Montfort UNiversity, Milton Keynes, May 1996, P123-135.

[36] B. Erol, F. Kossentini, *Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain*, IEEE trans. on Multimedia, vol. 2, no 2, pp. 129-138, 2000

[37] M. Flickner et al., *Query by Image and Video Content : The QBIC System*, IEEE Computer, Vol. 28, No. 9, pp. 23-32, 1995.

[38] J.H. Friedman, *Exploratory projection pursuit*, J. Amer. Statist. Assoc., 82,249-266, 1987.

[39] T. Fung, *An Assessment of TM Imagery for Land-Cover Change Detection*, IEEE Trans. on Geoscience and Remote Sensing, 28(4) :681-684, 1990.

[40] George et McCulloch, *Approaches for Bayesian variable selection*, Statistica Sinica 7, 1997.

[41] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 6, 721-741, 1984

[42] N.J. Gordon, D.J. Salmond, A.F.M Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, IEE Proc.-F, vol. 140, no. 2,pp. 107-113, 1993.

[43] M.H. Hansen et B. Yu, *Model Selection and the Principle of Minimum Description Length*, J. American Statistical Association, vol 96, pp.746-774, 2001.

[44] P. Héas, M. Datcu, A.Giros, *Image Time Series Mining For Dynamic Scene Understanding*, Second Workshop on Temporal Data Mining in conjunction with the International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CANADA , July 23 2002.

[45] P. Héas, M. Datcu, A.Giros, *Trajectory of dynamic clusters in image time-series*, Analysis of Multi-Temporal Remote Sensing Images, Proceedings of Multitemp 2003, Editors : P. Smits and L. Bruzzone, Series in Remote Sensing, Vol. 3, pp 39-49, July 2003a

[46] P. Héas, M. Datcu, A.Giros, P.Marthon, *Image Time Series Mining For Dynamic Scene Understanding*, IEEE International geoscience and remote sensing symposium (IGARSS03), Toulouse, France, 2003b.

[47] P. Héas et M. Datcu, *Bayesian Learning on Graphs for Reasoning on Image Time-Series*, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Editors : R. Fisher, R. Preuss, U. von Tousaint, American Institute of Physics, AIP Conf. Proc., ISBN 0-7354-0217-5, ISSN 0094-243X, Vol. 735, pp. 127-137, 2004a.

[48] P. Héas, M. Datcu, A.Giros et P.Marthon, *Mining Image Time-Series*, IEEE International geoscience and remote sensing symposium (IGARSS04), Anchorage, Alaska, 2004b.

[49] P. Héas et M. Datcu, *Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning*, accepted for publication in IEEE transactions on Geoscience and Remote Sensing, 2004c.

[50] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, In Learning in Graphical Models, M. Jordan, ed.. MIT Press, Cambridge, MA, 1999.

[51] A Heinrichs, D Koubaroulis, B Levienaise-Obadia, P. Rovida, et J.M. Jolion, *Image indexing and content-based search using pre-attentive similarities*. In 6th RIAO Conference, PARIS, April 12-14 2000.

[52] T. Hofmann,*Probabilistic Latent Semantic Indexing*, Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, pp 50-57, 1999.

[53] P.J. Huber, *Projection pursuit*, Annals of Statistics, Vol. 13, No. 2, pp. 435-475, 1993.

[54] A. Inselberg,*The Plane with Parallel Coordinates, Special Issue on Computational Geometry*, The Visual Computer, Vol. 1, pp. 69-97, 1985.

[55] E.T. Jaynes, *Information Theory and Statistical Mechanics*, Phys. Rev., 106, 620, 1957

[56] E.T. Jaynes, *Prior Probabilities*, IEEE Transactions on Systems Science and Cybernetics, volume SSC-4, pages 227-241, 1968

[57] E.T. Jaynes, *Bayesian spectrum and chirp analysis*, Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems, C.R. Smith and G.J. Erickson eds., page 1, 1987.

[58] E.T. Jaynes, *Probability Theory - The Logic of Science*, G.L. Bretthorst (Ed.). Cambridge University Press, 2003.

[59] G. Jumarie, *Relative Information, theories and applications.* Editor : Hermann Haken. Publishers / Tokyo, 1990.

[60] H. Jeffreys,*Theory of Probability*, Oxford Univ. press, 1939.

[61] R.E. Kalman,*A new approach to linear filtering and prediction problems*, trans. Am.Soc.Mech.Eng., Series D, Journal of Basic Engineering, 82, 35-45.

[62] D. A. Keim and H. P. Kreigel,*Visualization Techniques for Mining Large Databases : A Camparison*, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996.

[63] E. Keogh, P. Smyth, *A Probabilistic Approach to Fast Pattern Matching in Time-Series Databases*, Proc. of Knowledge Discovery in Databases and Data Mining, pp. 24-30, New Port Beach, USA, 1997.

[64] E. Keogh, M. Pazzani, *An Enhanced Representation of Time-Series which Allows Fast and Accurate Classification, Clustering and Relevance Feedback*, Knowledge Discovery in Databases and Data Mining, pp. 239-241, Madison, USA, 1998.

[65] A. Khokhar, Y.F. Day, A. Ghafoor, *A Framework for Semantic Modeling of Video Data for Content-Based Indexing and Retrieval*, ACM Multimedia, 1999.

[66] S. Kirkpatrick, CD Gelatt Jr., MP Vecchi, *Optimization by Simulated Annealing*,Science, 220, 4598, 671-680, 1983.

[67] K. Koffka, *Principles of Gestalt psychology*, New York : Harcourt, Brace, 1935.

[68] A.N. Kolmogorov,*Three Approaches to the quantitive definition of Information*, Problems of Information Transmission, 1, 1-17, 1965.

[69] S.Kullback. *Information theory and statistics*, Dover publications, inc. Mineola, New York, 1966.

[70] C. Lauvernet, F. Baret, F.X. Ledimet, *Assimilation of high temporal frequency Spot data to describe canopy functioning. The case of wheat crops in the ADAM experiment in Romania*,Proc. IGARSS, Toulouse, France, 2003.

[71] T.W. Lee, M. Girolami, A. J. Bell, T.J. Sejnowski, *A Unifying Information-theoretic Framework for Independent Component Analysis*, computers and mathematics with applications, Vol 31 (11), 1-21, March 2000.

[72] T.W. Lee, M.S. Lewicki, T.J. Sejnowski, *ICA Mixture Models for Unsupervised Classification of Non-Gaussian Sources and Automatic Context Switching in Blind Signal Separation*, IEEE Transactions on Pattern Recognition and Machine Intelligence 22(10), 1-12, 2000b.

[73] J. Li et J.Z. Wang, *Automatic Linguistic Indexing of Pictures by a statistical modeling approach*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp 1075-1088, 2003.

[74] T. Lin, H.J. Zhang, Q.Y. Shi, *Video Content Representation for Shot Retrieval and Scene Extraction*, International Journal of Image and Graphics, Vol. 1, No.3, 2001

[75] X.M. Liu, T. Chen, *Shot Boundary Detection Using Temporal Statistics Modeling*, IEEE Conf. on Acoustic, Speech and Signal Processing, ICASSP, Orlando, FL, U.S., 2002.

[76] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press. September 2003.

[77] D.J.C. MacKay, *Bayesian Modeling and Neural Networks*, PhD thesis, Dept. of Computation and Neural Systems, CalTech, 1992.

[78] P. Martin-Löf, *The Definition of Random Sequences*, Information and Control 9(6) : 602-619, 1966

[79] G. Miller, *The Magical Number Seven, Plus or Minus Two : Some Limits on Our Capacity for Processing Information*, Psycological Review 63 :2, pp. 81-96, 1956.

[80] Mohammad-Djafari A., *Entropie en traitement du signal*, Traitement du Signal, Vol. 11, No. 2, pp : 87-116, 1994

[81] M.G. Morton, *A computer Oriented Geodetic Data Base and a New Technique in File Sequencing*, IBM Ltd. Ottawa, Canada, 1996.

[82] M.R. Naphade, T. Kristjansson, B. Frey, T.S. Huang, *Probabilistic Multimedia Objects Multijects : A novel Approach to Indexing and Retrieval in Multimedia Systems*, Proc. IEEE International Conference on Image Processing, Volume 3, pages 536-540, Chicago, IL, 1998.

[83] M.R.Naphade, I.V. Kozintsev, T.S. Huang, *A Factor Graph Framework for Semantic Video Indexing*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 1, 2002.

[84] National American Space Agency
http ://topex-www.jpl.nasa.gov/science/time-series-data.html

[85] T.Niculiu et S. D. Cotofana, *Hierarchical Intelligent Simulation*, Proc. of the European Simulation Multiconference, pp. 243-246, 2001.

[86] C. Posse, *Analyse exploratrice de données et discrimination à l'aide de projection pursuit*, thèse 1124, école polytechnique fédérale de Lausanne, 1993.

[87] M.Quartulli et M.Datcu, *Stochastic geometrical modeling for urban scene understanding from a single sar intensity image with meter resolution*, IEEE Transactions on Geoscience and Remote Sensing,,41(9) :1976-1985, 2003

[88] J. Rissanen, *Modelingby shortest data description*,Automatica, vol 14, pp 465-471, 1978.

[89] J. Rissanen, *Stochastic complexity and modeling*, Ann. Statist. : 14, 1080-110, 1986.

[90] J. Rissanen, *Lectures on Statistical Modeling Theory*, Helsinki Institute for Information Technology, Finland, 2002.

[91] Y. Sakamoto, M.Ishiguro and G. Kitagawa, *Akaike information criterion statistics*. KTK Scientific Publishers / Tokyo, 1986.

[92] S. Santini et R. Jain, *Similarity is a geometer*, Multimedia Tools and Applications, vol. 5, no. 3, 1997.

[93] M. Schroeder, H. Rehrauder, K. Seidel et M. Datcu, *Spatial Information Retrieval from Remote Sensing Images - Part 2 : Gibbs Markov fields*, IEEE Transactions on Geoscience and Remote Sensing, 36 (5) :1446-1455, 1998.

[94] M. Schroeder, *WWW Online Demonstration : Interactive Learning and Probabilistic Retrieval.*,
http ://www.vision.ee.ethz.ch/ rsia/ClickBayes, 1999.

[95] M. Schroeder, H. Rehrauder, K. Seidel et M. Datcu, *Interactive learning and probabilistic retrieval in remoyte sensing image archives*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, pp. 2288-2298, 2000.

[96] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. :6 , 461-464, 1978.

[97] C.E. Shannon, *The Mathematical Theory of Communication*, The Univ. of Illinois Press, 1949.

[98] P.C. Smits, Alessandro Annoni, *Towards Specification-Driven Change Detection*, IEEE Trans. on Geoscience and Remote Sensing, 31(1) :136-145, 1993.

[99] R. Solomonoff, *A Formal Theory of Inductive Inference I and II*, Information and Control, 7, 1-22 and 224-254, 1964.

[100] A. Spartaru, *Fondements de la théorie de la transmission de l'information*, Lausanne : Presse Polytechniques Romandes, 1987

[101] C. Taskiran, E.J. Delp, C.A. Bouman, *The ViBE Video Database System : An Update and Further Studies*, Proc. of SPIE/I&ST, pp. 199-207, 2000.

[102] F. Topsoe, *Game theoritical equilibrium, maximum entropy and minimum information discrimination*, In a. Mohammad-Djafari and G.Demoments, editors, Maximum Entropy and Bayesian Methods, pages 15-23.Kluwer Academic Publishers, Dordrecht, Boston, London, 1993.

[103] D.A. Tran, K.A. Hua, K. Vu, *Semantic Reasoning based Video Database Systems*, Proc. of the 11th Int'l Conf. on Database and Expert Systems Applications, pp. 41-50, September 4-8, London, England, 2000.

[104] F. Tupin, H. Maître, J.-F. Mangin, J.-M. Nicolas, et E. Pechersky, *Detection of linear features in SAR images : application to road network extraction*, IEEE Trans. Geosci. and Remote Sensing vol. 36, no. 2, pp.434-453, mars 1998

[105] Vitanyi P., Li M., *Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity*, IEEE Trans. Inform. Theory, 46, 2000.

[106] C.S. Wallace et D.L. Boulton, *An information measure for classification*, Computing Journal 11 : 185-195, 1968.

[107] C.S. Wallace et D.L. Dowe, *Intrinsic Classification by MML - the Snob Program*, proc. of the 7th Australian Joint Conf. on Artificial Intelligence, pp 37-44, 1994.

[108] C.S. Wallace et D.L. Dowe, *Minimum Message Length and Kolmogorov Complexity*, special issue on Kolmogorov complexity, 42(4) : 270-283, 1999.

[109] F. Wang, *A Knowledge-Based Vision System for Detection Land Changes at Urban Fringes*, IEEE Trans. on Geoscience and Remote Sensing, 31(1) :136-145, 1993.

[110] K Warwick, P D Fox, E L Sutanto, *On merging gradient estimation with mean-tracking techniques for cluster identification*, Proc. 2nd IEEE Workshop on Computer Intensive Methods in Control and Signal Processing, Prague, Czech Republic, pp. 49-54, 1996.

[111] N. Wiener, *Cybernetics*, the MIT Press and J. Wiley and Sons, INC., New York. London, 1961.

[112] Y.F. Wong, *A New Clustering Algorithm Applicable to Multispectral and Polarimetric SAR Images*, IEEE Transactions on Geoscience and Remote Sensing, Vol.31, No.3, 1993.

[113] H. Zhang, J. Wu, D. Zhong, S.W. Smoliar, *An integrated system for content-based video retrieval ans browsing*, Pattern Recognition, Vol. 30, no. 4, pp. 643-658, 1997.

Remerciements

Je ne pourrais pas commencer les remerciements sans évoquer mon co-directeur de thèse, Mihai Datcu, qui a su guider et encourager mes recherches tout au long de ma thèse. Durant ces trois années, il m'a exprimé un réel soutient théorique de part sa profonde compréhension des méthodes stochastiques de traitement de l'information et sa pédagogie, mais aussi moral de part son énergie stimulante et sa confiance en moi. Je considère comme une chance le fait d'avoir pu travailler avec un superviseur et un chercheur aussi compétent.

Je souhaite aussi remercier Philippe Marthon, co-directeur de ma thèse, pour son amabilité et sa confiance. Je le remercie par ailleurs pour ses remarques et conseils qui m'ont permis d'améliorer la qualité de mon rapport.

Mais ce travail, et plus particulièrement la collaboration franco-allemande, entre le DLR, le CNES et l'IRIT, n'aurait pu avoir lieu sans la participation de Alain Giros. Cette expérience européenne m'a énormément apporté, autant sur le plan professionnel que personnel, et je tiens d'abord à exprimer mes remerciements pour son soutient dans cette aventure enrichissante. Je souhaite aussi le remercier pour les discussions stimulantes et ses critiques pertinentes qui m'ont permis de prendre un certain recul sur mes travaux. Finalement, j'aimerais le remercier pour m'avoir aidé à structurer mes idées, à développer une méthodologie efficace et une bonne démarche scientifique.

Je remercie également Henri Maître et Ali Mohammad-Djafari pour leur lecture attentive de mon manuscrit de thèse, leurs remarques judicieuses et leurs amabilités. Je remercie également Manuel Samuelides qui a accepté de faire partie du jury de thèse.

Ces quatre années de travail et d'échanges m'ont aussi permis d'enrichir ma culture dans de nombreux domaines et je ne peux que remercier Marco Quartulli et Ciro Delia pour m'avoir initier à la programmation orientée objet, Jordi Inglada pour ses idées et remarques originales, Hélène De Boissezon, Selma Cherchali et Aurélie Bouillon pour leur disponibilité et leur gentillesse, Andrea Pelizzari pour son aide dans la résolution de problèmes informatiques et Herbert Daschiel et Mariana Ciucu pour leur aide au travers de discussions enrichissantes.

Mais je tiens à ne pas oublier mes amis et ma famille qui ont fait que ces trois années de

thèse aient été très agréables. Je souhaite remercier vraiment spécialement Emilie pour sa patiente, son aide et sa compréhension durant cette dernière année de thèse parfois difficile. Je voudrais aussi spécialement remercier mes parents et mes frères pour leur amour et leur support. Je tiens aussi à spécialement remercier Anne-Sophie pour son soutient. Mais, je ne pourrais pas non plus oublier de remercier Xavier et David pour leur écoute et leurs conseils, Laurent pour sa bonne humeur, Jean pour m'avoir dissuader de déposer un brevet, Guillaume pour m'avoir apporté de l'air frais des Pyrénées, Samuel pour avoir toujours partager son repas, Marco pour ses chansons, Chris, Markus, Marian et Félix pour leur amitié et pour m'avoir fait découvrir la culture Allemande, Jean-Luc pour les discussions scientifiques autour d'un verre de rhum-citron, Loic pour son anti-militarisme, François-Xavier pour sa grande philosophie, Nicolas et Mathéo, Tibault, Michael, Stephane, Sylvain, Elen, Emmanuelle, José-Annibal, Olivier, Loic-San, Juan-Luis, Julian, Anaïs, Laëtitia ainsi que tout les amis avec qui j'ai passé ces trois années et, qui m'ont apporté par quelque façon un soutient dans cette entreprise qu'est la thèse.

Curriculum Vitae

Patrick HEAS
Né le 15 juin 1978 à Nantes, France

EXPERIENCE PROFESSIONNELLE

2004-2005 Centre National d'Études Spatiales (CNES), Toulouse. 5 mois
Réalisation d'un logiciel de fouille de séries temporelles d'images pour le projet ESA "Knowledge-centred Earth Observation"

2003-2004 Institut de Recherche en Informatique de Toulouse (IRIT). 18 mois
(thèse)

2001-2003 Centre aérospatial allemand (DLR), Allemagne. 18 mois
(thèse)

2001 Centre d'Étude Spatiale et du Rayonnement (CESR), Toulouse. 5 mois
(DEA, Sujet : Développement d'un système de segmentation d'images astronomiques par morphologie mathématique)

2001 Université Paul Sabatier (UPS), Toulouse. 5 mois
Enseignant vacataire du langage de programmation Caml

FORMATION

2001-2005 Doctorat de l'Ecole Nationale Supérieure de l'Aéronautique et de l'espace (SUPAERO), en informatique et télécommunication, Toulouse

2000-2001 Master de l'Ecole Nationale Supérieure de l'Aéronautique et de l'espace (SUPAERO) en signal, image et acoustique, Toulouse

1998-2000 Licence et maîtrise de Physique de l'université de Nantes