Fouille de données à partir de séries temporelles d'images satellites

Lynda Khiali

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale I2S

Unité de recherche TETIS

Fouille de données à partir de séries temporelles d'images satellites

Présentée par Lynda Khiali
Le 28/11/2018

Sous la direction de Maguelonne Teisseire
et Dino Ienco

Devant le jury composé de

Germain Forestier, Professeur, ENSISA, Université de Haute-Alsace	Rapporteur
Nicolas Méger, Maître de conférence HDR, LISTIC, IUT Annecy, Université Savoie Mont Blanc	Rapporteur
Carmen Gervet, Professeur, Espace-Dev, Université de Montpellier	Examinatrice
Camille Kurtz, Maître de conférence, LIPADE, Université de Paris Descartes	Examinateur
Maguelonne Teisseire, DR, TETIS, Irstea	Directrice
Dino Ienco, Chercheur HDR, TETIS, Irstea	Co-directeur

UNIVERSITÉ DE MONTPELLIER

Remerciement

En premier lieu, je tiens à remercier mes encadrants Maguelonne TEISSEIRE et Dino IENCO d'avoir accepté de diriger cette thèse et de m'avoir fourni d'excellentes conditions de travail. Je tiens à les remercier particulièrement pour leur disponibilité et leur investissement durant ces trois années, même dans les moments difficiles.

Je voudrais remercier également les rapporteurs de ma thèse : Prof. Germain Forestier et MCF. Nicolas Méger, ainsi que les membres du jury de ma thèse : Prof. Carmen Gervet et MCF. Camille Kurtz qui ont accepté d'évaluer mon travail. Merci pour leur intérêt, le temps consacré à la lecture de mon manuscrit ainsi que leurs questions et commentaires pertinents.

Je remercie tous mes amis et collègues de TETIS et Espace-Dev pour leur accueil et leur soutien.

Je remercie également mes amies Sakina, Asma, Anfel, Soumia, Nadira, Zineb et Samiha avec qui j'ai partagé de très beaux moments mais aussi des moments de stress durant ces trois années de thèse.

Je remercie particulièrement ma chère Sarah qui a toujours su trouver les bons mots pour me motiver, avec qui j'ai partagé autant les bons moments que ceux difficiles.

Je tiens à remercier du plus profond de mon cœur mes parents Youcef KHIALI et Aldjia ASMA, mes soeurs Souad et Sandra et mes frères Rabah et Rayanne. Merci pour votre soutien et vos encouragements. Je vous aime.

Résumé

Les images satellites représentent de nos jours une source d'information incontournable. Elles sont exploitées dans diverses applications, telles que : la gestion des risques, l'aménagent des territoires, la cartographie du sol ainsi qu'une multitude d'autre taches. Nous exploitons dans cette thèse les Séries Temporelles d'Images Satellites (STIS) pour le suivi des évolutions des habitats naturels et semi-naturels. L'objectif est d'identifier, organiser et mettre en évidence des patrons d'évolution caractéristiques de ces zones.

Nous proposons des méthodes d'analyse de STIS orientée objets, en opposition aux approches par pixel, qui exploitent des images satellites segmentées. Nous identifions d'abord les profils d'évolution des objets de la série. Ensuite, nous analysons ces profils en utilisant des méthodes d'apprentissage automatique. Afin d'identifier les profils d'évolution, nous explorons les objets de la série pour déterminer un sous-ensemble d'objets d'intérêt (entités spatio-temporelles/objets de référence). L'évolution de ces entités spatio-temporelles est ensuite illustrée en utilisant des graphes d'évolution.

Afin d'analyser les graphes d'évolution, nous avons proposé trois contributions. La première contribution explore des STIS annuelles. Elle permet d'analyser les graphes d'évolution en utilisant des algorithmes de clustering, afin de regrouper les entités spatio-temporelles évoluant similairement. Dans la deuxième contribution, nous proposons une méthode d'analyse **pluriannuelle** et **multi-site**. Nous explorons plusieurs sites d'étude qui sont décrits par des STIS pluriannuelles. Nous utilisons des algorithmes de clustering afin d'identifier des similarités intra et inter-site. Dans la troisième contribution, nous introduisons une méthode d'analyse semi-supervisée basée sur du clustering par contraintes. Nous proposons une méthode de sélection de contraintes. Ces contraintes sont utilisées pour guider le processus de clustering et adapter le partitionnement aux besoins de l'utilisateur.

Nous avons évalué nos travaux sur différents sites d'étude. Les résultats obtenus ont permis d'identifier des profils d'évolution types sur chaque site d'étude. En outre, nous avons aussi identifié des évolutions caractéristiques communes à plusieurs sites. Par ailleurs, la sélection de contraintes pour l'apprentissage semi-supervisé a permis d'identifier des entités profitables à l'algorithme de clustering. Ainsi, les partitionnements obtenus en utilisant l'apprentissage non supervisé ont été améliorés et adaptés aux besoins de l'utilisateur.

Table des matières

Table des figures

Liste des tableaux

Notations et préliminaires

Chapitres 4 et 5

I	Image satellite
l	Indice des dates d'acquisition
K	Nombre d'images satellites
T_l	Date d'acquisition
I_{T_l}	Image satellite acquisse en temps T_l
o	Objet
i, j	Indices des objets
$o^i_{T_l}$	Objet numéro i identifié dans l'image I_{T_l}
O	Ensemble d'objet
O_{T_l}	Ensemble d'objets identifiés dans l'image I_{T_l}
$\mathcal{O}$	Union de l'ensemble d'objets identifiés dans toutes les images de la série temporelle
$Pix(o)$	Ensemble des pixels de l'objet o
$Info(o)$	Vecteur de valeurs radiométriques de l'objet o
o^*	Objet de référence
G_{o^*}	Graphe d'évolution
V_{o^*}	Ensemble de nœuds du graphe d'évolution
E_{o^*}	Ensemble d'arcs du graphe d'évolution
syn	Synopsis
$\widetilde{O}$	Objet du synopsis
$\widetilde{O_{T_l}}$	Objet du synopsis correspondant à l'image I_{T_l}
$syn[T_l]$	Objet $\widetilde{O}_{T_l}$
α, σ_1, σ_2	Paramètres

Chapitres 2 et 6

$\mathbb{E}$	Ensemble de données
n	Cardinalité de l'ensemble $\mathbb{E}$
i, i', i''	Indices des entités de données
E_i	Entité numéro i de l'ensemble $\mathbb{E}$
$Dist(E_i, E_{i'})$	Distance entre l'entité E_i et $E_{i'}$
d	Dimension de l'espace de données
k	Nombre de cluster
k'	Nombre de classe
j, j'	Indices des clusters, centre des clusters, classes et partitionnements
C_j	Cluster numéro j
Cl_j	Classe numéro j
μ	Ensemble des centres de tous les clusters
μ_j	Centre du cluster C_j
L_{E_i}	Label de l'entité E_i
$\mathbb{P}$	Ensemble de partionnements
φ	Partitionnement expert
P	Partitionnement
P_j	Partitionnement numéro j
$L_{E_i}^{P_j}$	Label de l'entité E_i dans le partitionnement P_j

Chapitre 1

Introduction

Sommaire

1.1 Contexte et Motivations

1.1.1 Contexte

L'étude et la compréhension de l'environnement ainsi que l'analyse de l'occupation du sol et son évolution sont des thématiques d'une importance incontournable. Elles contribuent à la prise de décision dans divers domaines tels que la gestion des risques et des ressources naturelles, la gestion des cultures, la conservation de la biodiversité, etc. Les images satellites constituent de nos jours une source importante pour l'étude de l'environnement.

Les images satellites permettent de décrire la surface de la terre, elles permettent aussi de décrire son évolution à travers des séries temporelles. L'information contenue dans ces images est la principale source pour la cartographie du sol ainsi que la détection des changements et des dynamiques des territoires.

L'acquisition d'images satellites a connu une effervescence, nous comptons aujourd'hui des centaines voir des milliers de satellites d'observations de la terre. Ces satellites permettent l'acquisition de grandes quantités d'images. Le stockage de ces grandes quantités de données n'est plus problématique. En effet l'avancement technologique a permis de faire évoluer les techniques de stockages. Certaines de ces données sont mises gratuitement à la disponibilité des futurs utilisateurs (notamment les images satellites Sentinel-2). Ainsi, la problématique ne concerne plus le stockage ou la mise à disponibilité des données mais leur traitement et leur analyse.

1.1.2 Problématique

Le traitement des images satellites est une tâche fastidieuse. En effet, les méthodes d'analyse manuelles sont robustes et fiables. Cependant, elles requièrent un expert qui connait la zone d'étude. En outre, elles demandent un temps et un effort considérable. Ainsi, il est nécessaire de développer des méthodes automatiques et semi-automatiques pour le traitement d'images satellites.

Les méthodes d'apprentissage automatique ont été exploitées pour le traitement d'images dans différents domaines. Ces méthodes permettent d'extraire des connaissances à partir des données. Nous distinguons trois familles de méthodes : supervisées, non supervisées et semi-supervisées. Ces méthodes visent à regrouper ou classer les données en se basant sur leurs caractéristiques.

Les méthodes d'apprentissage supervisé se basent sur des données étiquetées permettant d'apprendre un modèle. Ce modèle est ensuite utilisé pour classer de nouvelles données. Les données étiquetées sont coûteuses et difficiles à générer. Comme alternative, nous exploitons les méthodes non supervisées. Ces méthodes ont l'avantage de ne pas utiliser de données étiquetées. Elles permettent de regrouper les données en groupes homogènes. Leur inconvénient est qu'elles présentent des résultats qui nécessitent du temps pour être interprétés. À mi-chemin de ces deux méthodes, on distingue les méthodes d'apprentissage semi-supervisé qui utilisent

des données étiquetées et des données non étiquetées. Ces méthodes sélectionnent un sous-ensemble de données à étiqueter. Elles identifient le plus petit sous-ensemble d'entités profitable à l'analyse de données. Le choix des entités à étiqueter est un point critique.

La littérature foisonne de travaux traitant de l'analyse d'images satellites via les méthodes d'apprentissage automatique que ce soit supervisé, semi-supervisé ou non supervisé. Ces travaux traitent de la plus petite unité de l'image qui est le pixel. Cependant les images satellites illustrent un paysage défini par différentes entités représentatives telles que les forêts, les parcelles agricoles, les zones urbaines, etc. La segmentation d'images satellites rend possible la détection de ces différentes entités du paysage et offre ainsi une nouvelle représentation de l'image beaucoup plus proche de la perception humaine.

Dans notre thèse, nous explorons l'analyse de Séries Temporelles d'Images Satellites (STIS) en utilisant des méthodes d'apprentissage. D'une part, nous comparons la représentation basée objet à la représentation basée pixel afin d'identifier la représentation la plus adéquate à l'analyse d'images satellites. D'une autre part, nous comparons les méthodes d'apprentissage non supervisé aux méthodes d'apprentissage semi-supervisé.

1.1.3 Objectif et contributions

Notre objectif est l'étude des dynamiques des habitats naturels et semi-naturels en exploitant des séries temporelles d'images satellites en utilisant des méthodes d'analyse orientée objet (OBIA- Object Oriented Image Analysis) combinées aux méthodes d'apprentissage automatique.

L'analyse basée objet des STIS implique la segmentation des images et l'illustration de l'évolution des objets résultats. Ces objets sont caractérisés par trois dimensions : spatiale, temporelle et spectrale. Il est nécessaire d'adopter une représentation tenant compte de ces trois dimensions.

Dans leur travaux, (Guttler et collab., 2017) proposent une représentation basée sur les graphes. Différemment des travaux de l'état de l'art, les images sont segmentées indépendamment. En effet, chaque image est caractérisée par son propre paysage et résulte en un ensemble d'objets différents. Nous avons adopté cette représentation afin d'illustrer l'évolution des différents objets d'intérêt à travers la série.

Dans notre première contribution, nous avons défini un système permettant d'identifier des entités spatio-temporelles dans des images satellites segmentées et de représenter leur évolution en construisant des graphes nommés graphes d'évolution. Ce système permet en outre d'analyser les graphes résultats en utilisant des algorithmes de clustering afin d'identifier des entités évoluant similairement et de mettre en évidence des patrons d'évolution. Cette méthode a été évaluée en utilisant des STIS annuelles.

Dans notre deuxième contribution, notre objectif est d'analyser des évolutions

à partir de séries temporelles d'images satellites pluri-annuelles. Nous avons adapté la représentation par graphe afin de pouvoir illustrer l'évolution des objets d'intérêt (entités spatio-temporelles) au travers de ces séries. Notre objectif est également de réaliser une analyse multi-site en regroupant plusieurs sites simultanément, afin d'identifier des entités spatio-temporelles dont les évolutions sont similaires intra et inter-sites. Pour regrouper ces entités, nous avons exploité des méthodes d'apprentissage non supervisé.

Dans la dernière contribution, nous avons exploité les méthodes d'apprentissage semi-supervisé dans l'analyse des séries temporelles d'images satellites. Notre objectif est de guider le processus de clustering et de l'adapter aux besoins de l'utilisateur. En effet, plusieurs partitionnement sont possibles pour un même ensemble de données selon la dimension considérée. Nous avons ainsi introduit des connaissances expertes dans le processus de clustering en étiquetant un sous-ensemble des données.

1.2 Organisation du mémoire

Cette thèse est organisée en sept chapitres. Le premier chapitre est introductif, il permet de présenter brièvement le contexte général de la thèse, ses problématiques et les différentes contributions réalisées.

Le deuxième et le troisième chapitre permettent de situer le contexte de la thèse, ils sont nécessaires pour la bonne compréhension de nos travaux.

Nos travaux de thèse exploitent des méthodes d'apprentissage automatique pour l'analyse des séries temporelles d'images satellites. Le chapitre 2 introduit les concepts nécessaires pour la compréhension de la thèse. Il est organisé en trois parties. Dans la première partie, nous définissons l'apprentissage automatique et nous présentons les trois méthodes d'apprentissage : supervisé , non supervisé et semi-supervisé. En outre, nos décrivons les algorithmes de clustering non supervisés utilisés dans nos travaux. La deuxième partie traite des images satellites, nous définissons les images satellites et leurs propriétés radiométriques. Puis, nous présentons leur processus d'acquisition. La troisième partie traite de l'analyse des séries temporelles d'images satellites par approche automatique. Nous présentons les différents travaux d'analyse basée pixel et d'analyse basée objet, notre objectif est de comparer les deux représentations.

Dans le chapitre 3 , nous décrivons nos zones d'étude. Nous avons évalué nos différentes contributions sur quatre zones d'étude. Nous présentons d'abord leurs localisation géographique et caractéristiques. Puis nous introduisons les séries temporelles d'images satellites utilisées. Nous présentons aussi les différentes étapes de leur pré-traitement comprenant la segmentation, caractérisation des objets par bandes spectrales et indices radiométriques. Enfin, nous présentons leurs cartes d'occupation du sol réalisées par un expert du terrain. Ces cartes sont utilisées pour valider nos résultats.

Les chapitres 4, 5 et 6 présentent nos trois contributions. Dans le chapitre 4 nous définissons une méthode d'analyse de séries temporelles d'images satellites annuelles ayant pour objectif le suivi de l'évolution des habitats naturels est semi-naturels (Khiali et collab., 2018). Le chapitre 5 introduit une méthode d'analyse de séries temporelles pluri-annuelles et multi-site (Khiali et collab., 2019).

Pour chacune de ces deux contributions, nous présentons la méthode générale puis nous décrivons chacune de ses étapes. Nous décrivons le processus de détection des objets d'intérêt (entités spatio-temporelles/objets de référence) à analyser. Pour chaque entité spatio-temporelle, nous identifions les objets lui correspondant, qui nous permettrons de construire les graphes d'évolution. Les graphes d'évolution sont transformés en synopsis qui résument l'information radiométrique contenue dans les graphes. Nous calculons ensuite la distance entre les synopsis en se basant sur des mesures adaptées aux spécificités de chaque série temporelle (annuelle ou pluri-annuelle). Enfin, nous appliquons un algorithme de clustering pour regrouper les entités spatio-temporelles évoluant similairement. Dans la première contribution, les entités de chaque site sont regroupées indépendamment. Tandis que dans la deuxième contribution, les entités identifiées sur chaque site sont regroupées simultanément afin d'identifier des similarités intra et inter-sites. Les deux contributions ont été validées sur des données réelles, nous décrivons le protocole expérimental adopté puis nous présentons et discutons les résultats obtenus.

Dans le chapitre 6, nous introduisons une méthode d'analyse de séries temporelles d'images satellites semi-supervisée. Notre objectif est d'améliorer les résultats obtenus en utilisant l'apprentissage semi-supervisé. L'apprentissage semi-supervisé exploite des données étiquetées. Nous présentons notre approche de sélection des entités à étiqueter basée sur le clustering par ensemble. Puis nous décrivons l'algorithme de clustering utilisé permettant de traiter conjointement les données étiquetées et les données non étiquetées. Enfin nous introduisons notre protocole expérimental et nous reportons les résultats obtenus.

Le dernier chapitre résume les différentes propositions réalisées au cours de notre thèse. En outre, nous identifions les différentes pistes permettant d'améliorer nos travaux actuels. Nous identifions des perspectives en continuité pour nos propositions.

1.3 Publications

Nous travaux de thèse ont donné lieu à plusieurs publications, que nous listons ci-dessous.

Revue internationale avec comité de lecture

- Khiali, L. ; Ienco, D. and Teisseire, M. "Object-oriented satellite image time series analysis using a graph-based representation", Ecological Informatics,

2018, vol. 43 p. 52-64, doi :10.1016/j.ecoinf.2017.11.003 , URL. https ://-doi.org/10.1016/j.ecoinf.2017.11.003.

- Khiali, L., M. Ndiath, S. Alleaume, D. Ienco, K. Ose et M. Teisseire. 2019, "Detection of spatio-temporal evolutions on multi-annual satellite image time series : A clustering based approach", International Journal of Applied Earth Observation and Geoinformation, vol. 74, p. 103-119, doi :10.1016/j.jag.2018.07.014 , URL. https ://-doi.org/10.1016/j.jag.2018.07.014.

Acte de conférence nationale avec comité de lecture

- Khiali, L., D. Ienco, et M. Teisseire (2017). "Analyse des dynamiques spatio-temporelles à partir de séries temporelles d'images satellitaires". In Extraction et Gestion des Connaissances, Volume E-33 of RNTI, pp. 261–272.

Communication internationale

- Khiali, L., M. Ndiath, S. Alleaume, D. Ienco, K. Ose et M. Teisseire. 2018, "Multi-annual Satellite Image Time Series Analysis using an Object-Oriented Approach", 7th GEOBIA conference.

Poster

- Khiali, L., D. Ienco, et M. Teisseire. 2017, "Object-Oriented Satellite Image Time Series Analysis through a Graph-Based representation", Journée Sciences des Données MaDICS 2017.

Chapitre 2

Apprentissage automatique et images satellites

Sommaire

2.1 Introduction

L'apprentissage automatique désigne le processus de fonctionnement d'un système d'intelligence artificielle basé sur l'apprentissage et le raisonnement. Les méthodes d'apprentissage automatique permettent d'analyser les données pour établir des corrélations entre les entités afin d'en extraire des connaissances. En outre, l'apprentissage automatique permet de pallier les problèmes de l'analyse de grandes quantités de données. Le traitement de ces données est fastidieux, d'où le besoin d'utiliser des méthodes automatiques.

L'apprentissage automatique permet d'analyser différents types de données tels que : les textes, les vidéos, les images, etc. Ces méthodes ont été largement exploitées pour le traitement d'images dans divers domaines tels que : la sécurité (reconnaissance faciale), la médecine (détection de tumeurs), l'astronomie (détection des mouvements des corps célestes) et la télédétection.

Dans le domaine de la télédétection, l'apprentissage automatique permet d'analyser des images satellites pour la cartographie du sol et d'analyser des séries temporelles pour la détection de changements. Dans cette thèse, nous exploitons les séries temporelles pour le suivi de l'évolution des habitats naturels et semi-naturels en utilisant des méthodes d'apprentissage automatique.

Dans ce chapitre, nous présentons les concepts nécessaires à la compréhension de nos travaux de thèse. Dans la première partie, nous définissons l'apprentissage automatique et ses différentes méthodes : supervisées, non supervisées et semi-supervisées (Section 2.2). Dans la deuxième partie, nous définissons les concepts de base de la télédétection : les images satellites, leur processus d'acquisition et leurs propriétés spectrales (Section 2.3). Dans la dernière partie, nous présentons les approches d'analyse des séries temporelles par apprentissage automatique de la littérature (Section 2.4).

2.2 Apprentissage automatique

L'Apprentissage Automatique (AA) ou Machine Learning (ML) en anglais est une branche de l'intelligence artificielle. Il est défini par Herbert Simon, comme étant des changements dans un système lui permettant de réaliser de meilleures performances lors de la répétition d'une même tâche sur la même population (Herbert, 1983). L'apprentissage automatique permet ainsi à un système de s'améliorer automatiquement avec l'expérience. Il permet en outre à un système d'apprendre une tache sans être explicitement programmé.

L'apprentissage automatique est aujourd'hui exploité dans divers domaines comprenant : la reconnaissance d'objets, la reconnaissance vocale, le traitement du langage naturel, l'analyse d'images, l'analyse financière, la bio-informatique etc.

Il existe dans la littérature plusieurs algorithmes permettant à un système d'apprendre. Ces algorithmes sont classés selon leur méthode d'apprentissage : les méthodes d'apprentissage supervisé, les méthodes d'apprentissage semi-supervisé et les

méthodes d'apprentissage non supervisé. Nous allons définir dans ce qui suit chacune de ces trois méthodes.

2.2.1 Méthode d'apprentissage supervisé

Les méthodes d'apprentissage supervisé appelées aussi méthodes de classification, permettent d'analyser des données afin d'attribuer à chaque entité un label (étiquette) correspondant à sa classe. Le principe de ces méthodes est d'établir un ensemble de règles sous forme de modèle permettant de classer les entités en se basant sur leurs attributs.

L'apprentissage supervisé est réalisé en deux étapes : la phase d'apprentissage et la phase de classification. La phase d'apprentissage permet d'apprendre le modèle de classification. La construction du modèle se base sur des données d'apprentissage étiquetées au préalable. La phase de test permet d'évaluer les performances du modèle et sa capacité à identifier les classes de nouvelles entités inconnues.

Il existe dans la littérature plusieurs méthodes de classification (Nasrabadi, 2007) dont les arbres de décision et les machines à vecteurs de support qui sont les plus populaires.

Les machines à vecteurs de support (Cortes et Vapnik, 1995) se basent sur l'utilisation de fonctions dites noyaux (kernel). Cette méthode permet de séparer linéairement les données en deux classes. Elle identifie deux hyperplans parallèles définissant une marge qui sépare les données. L'objectif est d'identifier la marge maximale. Les machines à vecteurs de support permettent aussi de gérer les données non séparables linéairement en changeant leur espace de description afin de rendre possible une séparation linéaire. Cette méthode permet aussi de séparer les données en plusieurs classes (nombre de classes supérieur à 2) en traitant chaque paire de classes indépendamment puis en combinant les résultats (Vapnik, 1998).

Les limites des méthodes supervisées consistent en leur besoin de données étiquetées.

2.2.2 Méthode d'apprentissage non supervisé

Les méthodes d'apprentissage non supervisé englobent des algorithmes d'exploration de données visant à inférer des connaissances de ces données. Le clustering de données appartient à la famille des approches non supervisées. Les algorithmes de clustering visent à décomposer un ensemble d'entités en plusieurs sous-ensembles les plus homogènes possible. Chaque ensemble comprend des entités similaires, tandis que les entités appartenant à des groupes différents sont dissimilaires (Kaufman et Rousseeuw, 2009). Afin de définir des groupes de données les algorithmes de clustering se basent sur des mesures de distance (ou de similarité). Le clustering compte quatre étapes comme illustré dans la Figure 2.1 (Xu et Wunsch, 2005). La première étape consiste à identifier les attributs caractéristiques des données à analyser. Les algorithmes de clustering utilisent ces attributs afin d'estimer la simi-

larité entre les entités et de générer un partitionnement des données. Les groupes de données générés sont ensuite examinés afin de vérifier la qualité du partitionnement en utilisant des mesures de qualité. La dernière étape consiste à interpréter le partitionnement généré afin d'en extraire des connaissances.

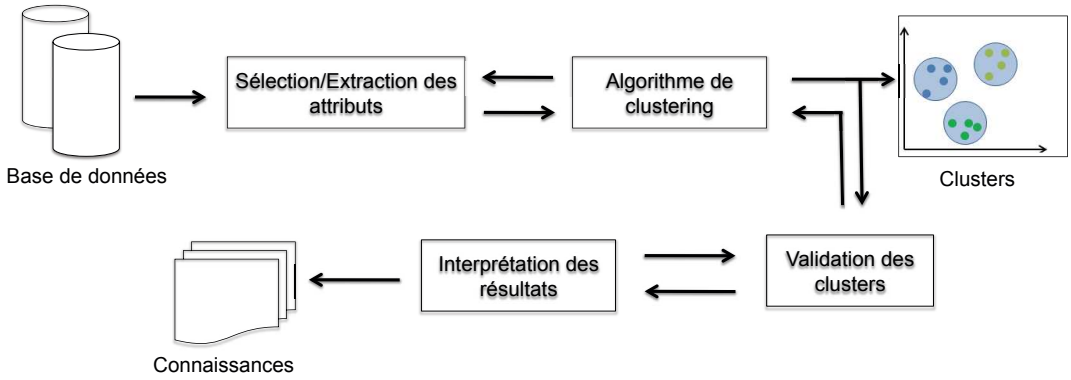


FIGURE 2.1 – Processus général du clustering des données.

Traditionnellement, les techniques de clustering sont classées en trois familles (RAI et SINGH, 2010) : méthodes hiérarchiques, méthodes de partitionnement et méthodes basées sur la densité. Nous comptons dans la littérature plus de neuf familles de méthodes (XU et TIAN, 2015) qui sont basées sur : les graphes, les probabilités, les modèles etc.

Nous présentons dans ce qui suit les deux types de méthodes exploitées dans cette thèse qui sont : les méthodes hiérarchiques et les méthodes de partitionnement

Méthodes hiérarchiques

Les méthodes de clustering hiérarchique comptent des algorithmes itératifs et récursifs. On distingue deux types : les algorithmes agglomératifs et les algorithmes diviseurs.

Les algorithmes agglomératifs forment initialement n clusters, chaque entité forme un cluster singleton. Les clusters les plus similaires sont ensuite fusionnés de manière itérative jusqu'à obtenir un seul cluster contenant toutes les entités. Tandis que les algorithmes diviseurs forment initialement un seul cluster contenant les n entités, ce cluster est ensuite divisé jusqu'à obtenir n clusters singletons. Les clusters sont représentés sous forme de dendrogramme. Un dendrogramme est défini souvent par un arbre binaire qui permet d'illustrer le regroupement des entités généré à chacune des itérations de l'algorithme. Le résultat final du clustering est obtenu en coupant le dendrogramme au niveau souhaité. La Figure 2.2 illustre un exemple de dendrogramme résultat du clustering agglomératif de cinq entités $\{E_1, E_2, E_3, E_4, E_5\}$.

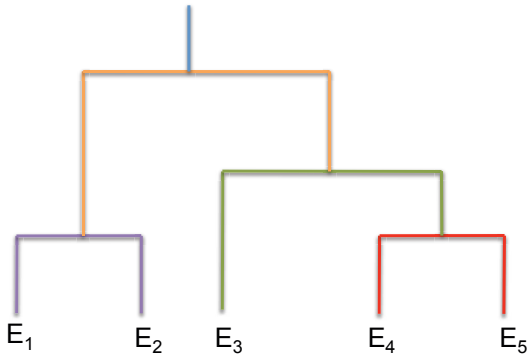


FIGURE 2.2 – Exemple de dendrogramme illustrant le clustering de cinq entités de données.

L'Algorithme 1 définit le clustering agglomératif (COHEN-ADDAD et collab., 2018). L'étape une consiste à créer n clusters singletons afin d'initialiser le dendrogramme. L'étape trois consiste à regrouper itérativement les deux clusters les plus similaires. Le dendogramme est ensuite mis à jour à chaque itération.

Le calcul de la distance entre deux clusters est défini dans l'étape deux de trois manières différentes. Le saut minimum définit la distance en retenant le minimum des distances entre leurs entités, tandis que le saut maximum définit la distance en retenant le maximum des distances entre leurs entités. Enfin, le saut moyen définit la distance en calculant la moyenne des distances entre leurs entités.

Algorithme 1 : Algorithme agglomératif du clustering hiérarchique

1: ENTRÉES : DONNÉES $\mathbb{E}= \{E_i\}_{i=1}^{n} \in \mathbb{R}^d$
2: SORTIES : *dendrogramme*
3: (1) Création de n clusters singleton : $P = \{C_1, C_2, .., C_n\}$
4: (2) Définir la distance $D(C_j, C_{j'}) = \begin{cases} \frac{1}{|C_j||C_{j'}|} \sum_{E_i \in C_j, E_{i'} \in C_{j'}} Dist(E_i, E_{i'}) \text{ Saut moyen} \\ argmin_{E_i \in C_j, E_{i'} \in C_{j'}} Dist(E_i, E_{i'}) \text{ Saut minimum} \\ argmax_{E_i \in C_j, E_{i'} \in C_{j'}} Dist(E_i, E_{i'}) \text{ Saut maximum} \end{cases}$
5: (3). Création du dendogramme
6: **pour** $k = n$ à 1 **faire**
7: $C_x, C_y = argmin_{C_j, C_{j'} \in P} D(C_j, C_{j'})$
8: $C_x = C_x \cup C_y$
9: *Suppression du cluster* C_y
10: *Mise à jour du dendogramme*
11: **fin pour**
12: **retourner** *dendogramme*

Méthodes de partitionnement

Différentes des méthodes hiérarchiques, les méthodes de partitionnement assignent les entités à k clusters différents d'une manière itérative mais non hiérarchique. En théorie le nombre de partitionnements possible est démesuré même pour un petite ensemble de données (XU et WUNSCH, 2005). Afin de diminuer la complexité du problème, les méthodes de partitionnement se basent sur des heuristiques. En supposant que le nombre de clusters est connu à priori, ces méthodes sélec-

tionnent (souvent de manière aléatoire) k entités de données qui sont utilisées pour initialiser les centres des clusters. Les entités de données sont affectées aux différents clusters en se basant sur leur proximité aux différents centres sélectionnés. Les centres ainsi que l'affectation des entités sont ensuite mis à jour itérativement afin d'améliorer le partitionnement initial.

La mise à jour des clusters se base sur une fonction objective. La fonction objectif permet de qualifier la qualité du clustering. L'objectif de ces méthodes est de minimiser leur fonction objectif (Gan et collab., 2007). La fonction d'erreur quadratique est l'une des fonctions objectives les plus utilisées dans la littérature.

$$J(P) = \sum_{j=1}^{k} \sum_{E_i \in C_j} \left\| (E_i - \mu_j) \right\|^2$$

Cette fonction est calculée en utilisant la somme de la distance euclidienne des entités au centre de leur cluster. Les algorithmes les plus populaires dans cette famille de méthodes sont le Kmeans (MacQueen et collab., 1967) et le Kmédoïdes (Kaufman et Rousseeuw, 1987) (Gan et collab., 2007). Les étapes du Kmeans sont décrites dans l'Algorithme 4.

Parmi les méthodes de partitionnement, nous comptons le clustering spectral. Cette méthode est basée sur l'algorithme Kmeans. À la différence de ce dernier, le clustering spectral redéfinit l'espace de représentation des données (réduction de dimension). Il exploite pour cela les vecteurs propres de la matrice du Laplacien de la matrice de similarité des données. Les étapes du clustering spectral sont décrites dans l'Algorithme 2 (Von Luxburg, 2007) :

Algorithme 2 : Algorithme Spectral

Entrées : Matrice de similarité S $\in \mathbb{R}^{n*n}$de l'ensemble de données $\mathbb{E}$
Nombre de clusters k
Sorties : $P = \{C_1, C_2, .., C_k\}$
(1). Calculer la matrice laplacienne L de la matrice S
(2). Calculer les premiers h vecteurs propres $v_1, v_2, v_3, ..., v_h$ de la matrice L
(3). Définir la matrice $V \in \mathbb{R}^{n*h}$ en assimilant les vecteurs propres aux colonnes
(4). Définir l'ensemble des données $\mathbb{E}$ dans le nouvel espace d'attribut :
E_i sera défini par les attributs de la ième ligne de la matrice V
(5). Clustering des données en utilisant le Kmeans
retourner $P = \{C_1, C_2, .., C_k\}$

2.2.3 Méthode d'apprentissage semi-supervisé

L'apprentissage semi-supervisé est un consensus entre l'apprentissage supervisé et l'apprentissage non supervisé. Il exploite à la fois des données étiquetées et des données non étiquetées. Il vise à utiliser le minimum possible de données étiquetées non redondantes qui permettent d'améliorer la qualité de regroupement. On distingue deux types de méthodes d'apprentissage semi-supervisé : (i) la classification semi-supervisée (ii) le clustering semi-supervisé.

La classification semi-supervisée permet l'ajout de données non étiquetées pour la construction du modèle. Tandis que pour le clustering semi-supervisé, on ajoute

des données étiquetées afin de guider le processus de clustering.

Des approches automatiques et semi-automatiques ont été proposées dans la littérature, elles permettent de sélectionner des données non étiquetées et des données étiquetées pertinentes pour la classification et le clustering respectivement.

Parmi les méthodes de sélection des entités non étiquetées pour la classification semi-supervisée, nous distinguons principalement les suivantes : l'auto-apprentissage, le co-apprentissage et l'apprentissage multi-vues.

L'auto-apprentissage est l'une des premières méthodes utilisées en classification semi-supervisée (Yarowsky, 1995). L'algorithme de classification utilise l'ensemble de données étiquetées pour apprendre un modèle initial, Ce modèle est ensuite utilisé afin de classer les données non étiquetées. Parmi les nouvelles données étiquetées, les plus confidentes sont sélectionnées. L'algorithme de classification réapprend un nouveau modèle en utilisant les données étiquetées initiales ainsi que les nouvelles données étiquetées par le modèle construit (Maulik et Chakraborty, 2011). Le co-apprentissage divise les descripteurs des données en deux sous-ensembles. L'algorithme de classification utilise les deux sous-ensembles pour apprendre deux modèles. Le premier modèle est ensuite utilisé pour classifier les données non étiquetées. Les données dont la classification est la plus fiable sont sélectionnées et ajoutées à l'ensemble de données étiquetées. Le deuxième modèle est réentraînement en utilisant les données étiquetées initiales ainsi que les nouvelles données étiquetées. Le même processus peut être réalisé en inversant les rôles des deux modèles. Ainsi, les données étiquetées par chacun des modèles sont utilisées pour réapprendre l'autre modèle (Zhu et collab., 2016). L'apprentissage multi-vues permet d'apprendre plusieurs modèles soit en utilisant plusieurs classificateurs (Roy et collab., 2014) soit en divisant l'ensemble d'attributs en plusieurs sous-ensembles (Sun, 2013). Les modèles sont ensuite utilisés pour classer les données. Les différentes classifications obtenues sont combinées en utilisant des méthodes de fusion, afin de générer une classification finale des données. Le vote majoritaire est l'une des méthodes de fusion les plus utilisées. Ainsi pour chaque donné, on attribue la classe prédite la plus fréquente.

Les méthodes de clustering semi-supervisé sont connues sous le nom de clustering par contraintes (Davidson et Basu, 2007). Les données étiquetées se présentent sous forme de contraintes. Les contraintes les plus utilisées sont définies sur une paire d'entités, l'expert identifie si elles sont similaires et doivent être regroupées ensemble (**Must-link**) ou bien si elles sont différentes et doivent être séparées (**Cannot-link**). Les entités sélectionnées sont celles pour lesquelles l'algorithme est le plus confus. Il existe dans la littérature plusieurs approches de sélection de contraintes (Abin, 2016 ; Abin et Beigy, 2014; 2015 ; Vu et collab., 2012). Dans (Abin et Beigy, 2014), les auteurs proposent une méthode permettant de représenter les données dans un espace puis d'identifier les différentes régions les décrivant. En considérant ces régions, ils identifient deux types de données ambiguës : les données proches mais appartenant à des régions différentes et les données distantes mais appartenant à la même région. En se basant sur ces hypothèses, cette méthode introduit une mesure d'utilité permettant d'estimer le potentiel des entités à améliorer le clustering.

L'apprentissage semi-supervisé peut être aussi réalisé d'une manière active (Settles, 2010). L'apprentissage actif est un processus itératif faisant intervenir un expert à chacune de ses étapes. La classification semi-supervisée active permet d'apprendre un modèle à partir d'un ensemble de données étiquetées, puis l'améliorer en ajoutant des données étiquetées par un expert à son jeu d'apprentissage. De même pour le clustering par contraintes actif, à chaque itération, on sélectionne des contraintes qui sont étiquetées par l'expert puis utilisées pour guider le clustering.

2.2.4 Indice de qualité de l'apprentissage automatique

L'évaluation des résultats d'une classification (apprentissage supervisé) correspond à vérifier si les différentes entités ont été étiquetées correctement ou pas par le modèle construit. Il existe dans la littérature plusieurs indices de qualité permettant d'évaluer les résultats et la robustesse des modèles de classification. Parmi les indices les plus utilisés, nous citons : la précision, le rappel et la F-mesure (Ferri et collab., 2009)

Cependant l'évaluation des résultats du clustering (apprentissage non supervisé) est une tache non triviale. Le clustering a pour objectif d'identifier des groupes d'entités cohérents au sein des données. Cependant, pour un même ensemble de données, il peut exister plusieurs partitionnements possibles. Comment peut-on identifier le meilleur partitionnement parmi ces différentes possibilités ? De nombreux travaux dans la littérature se sont intéressés à cette problématique (Arbelaitz et collab., 2013 ; Hämäläinen et collab., 2017 ; Rendón et collab., 2011).

Afin d'évaluer les résultats du clustering, on compte principalement deux approches : évaluation interne et évaluation externe. Ces approches sont détaillées dans les deux prochaines sections.

2.2.4.1 Indices de qualité internes

Les indices de qualité internes se basent sur le calcul de la cohésion et la séparabilité des clusters identifiés. La cohésion permet d'estimer la similarité des entités regroupées dans les clusters. La séparabilité des clusters permet d'estimer la différence entre les entités regroupées dans des clusters différents. Les valeurs de la cohésion et de la séparabilité sont ensuite combinées par une somme ou un ratio afin d'évaluer la qualité du clustering.

Il existe dans la littérature plusieurs indices de qualité internes dont : l'indice de Silhouette, l'indice de Davies-Bouldin (DB) et l'indice de Calinski-Harabasz (CH).

— Indice de Davies-Bouldin : L'indice de DB (Davies et Bouldin, 1979) estime la cohésion en calculant une somme normalisée de la distance des entités de chaque cluster à leur centre. Il estime la séparabilité en calculant les distances entre les centres des clusters. La solution de partitionnement qui minimise cette indice correspond au meilleur partitionnement. L'indice de DB est défini par l'Équation 2.1.

$$DB(P) = \frac{1}{k} \sum_{C_j \in P} argmax_{C_{j'} \in P \setminus C_j} \frac{S(C_j) + S(C_{j'})}{Dist(\mu_j, \mu_{j'})} \tag{2.1}$$

Où :
$S(C_j) = 1/|C_j| \sum_{E_i \in C_j} Dist(E_i, \mu_j)$

— Indice de Calinski-Harabasz : L'indice CH (Caliński et Harabasz, 1974) calcule un ratio de la séparabilité et la cohésion du partitionnement. La cohésion d'un cluster est estimée en calculant la distance de ses entités à leur centre. La séparabilité quant à elle est estimée en calculant la distance entre les centres des clusters et le centre de l'ensemble des données. Le partitionnement qui maximise la valeur de l'indice CH est le meilleur partitionnement. Cette indice est défini par l'Équation 2.2.

$$CH(P) = \frac{|\mathbb{E}| - k}{k - 1} * \frac{\sum_{C_j \in P} |C_j| Dist(\mu_j, \bar{E})}{\sum_{C_j \in P} \sum_{E_i \in C_j} Dist(E_i, \mu_j)} \tag{2.2}$$

Où :
$\bar{E}$ représente le centre de l'ensemble de données.

— Indice de Silhouette (Rousseeuw, 1987) : L'indice de silhouette calcule la somme de la séparabilité et la cohésion d'un partitionnement. Le partitionnent qui maximise la valeur de cet indice est le meilleur partitionnement. La cohésion d'un cluster est estimée en calculant la distance entre les entités du même cluster. La séparabilité est estimée en calculant la distance entre chaque entité et son voisin le plus proche appartenant à un autre cluster. Cette indice est défini par l'Équation 2.3.

$$Sil(P) = 1/|\mathbb{E}| \sum_{C_j \in P} \sum_{E_i \in C_j} \frac{b(E_i, C_j) - a(E_i, C_j)}{argmax\{a(E_i, C_j), b(E_i, C_j)\}} \tag{2.3}$$

Où :
$a(E_i, C_j) = 1/|C_j| \sum_{E_{i'} \in C_j} Dist(E_i, E_{i'})$
$b(E_i, C_j) = argmin_{C_{j'} \in P | C_j} 1/|C_{j'}| \sum_{E_{i'} \in C_{j'}} Dist(E_i, E_{i'})$

2.2.4.2 Indices de qualité externes

Afin de calculer des indices de qualité externes on se base sur une classification experte des données. Les deux partitionnements, le partitionnement automatique réalisé par un algorithme de clustering et le partitionnements expert sont comparés afin d'estimer leur similarité. Plus proche est le partitionnement automatique au partitionnement expert, meilleure est la qualité du clustering. Il existe dans la littérature plusieurs indices de qualité externes dont : l'Indice de Rand Ajusté, l'Information Mutuelle Normalisée et la Pureté.

— Pureté : Afin de calculer la pureté d'un partitionnement, on calcule d'abord la pureté de chaque cluster. On identifie pour chaque cluster la classe maximale de ces entités. La pureté d'un cluster est défini par l'équation 2.4.

$$Pureté(C_j) = \frac{1}{|C_j|}\ argmax_{Cl_{j'} \in \varphi} |C_j \cap Cl_{j'}| \tag{2.4}$$

La pureté du partitionnement est ensuite calculée en sommant la pureté des différents clusters ainsi que défini par l'équation 2.5. Cette indice varie dans un intervalle de $[0, 1]$, la valeur 1 indique une correspondance parfaite entre les deux partitionnements.

$$Pureté(P, \varphi) = \sum_{j=1}^{k} \frac{|C_j|}{|\mathbb{E}|} Pureté\ (C_j) \tag{2.5}$$

— L'Information Mutuelle Normalisée : L'Information Mutuelle Normalisé (NMI)(Thomas, 1991) calcule l'information partagée par les deux partitionnements, le partitionnement réalisé par un l'algorithme de clustering et la classification experte. Cette indice est défini, par l'équation 2.6 :

$$NMI(P, \varphi) = \frac{I(P, \varphi)}{\sqrt{H(P).H(\varphi)}} \tag{2.6}$$

Où :
$I(P, \varphi) = \sum_{C_j \in P} \sum_{Cl_{j'} \in \varphi} \frac{|C_j \cap Cl_{j'}|}{|\mathbb{E}|} \log \frac{|\mathbb{E}||C_j \cap Cl_{j'}|}{|C_j||Cl_{j'}|}$
$H(P) = -\sum_{C_j \in P} \frac{|C_j|}{|\mathbb{E}|} \log \frac{|C_j|}{|\mathbb{E}|}$

L'information mutuelle normalisée considère chaque paire de cluster et de classe. L'information partagée pour chaque paire est illustrée en utilisant une matrice nommée, Matrice de Confusion (MC). Cette Matrice, $MC(P, \varphi)$ est de taille, $|P| * |\varphi|$, ou chaque élément $m_{jj'}$ dénote le nombre d'entités en commun entre le cluster C_j et la classe $Cl_{j'}$, comme illustré par la Table 2.1.

$P \setminus \varphi$	Cl_1	Cl_2	...	$Cl_{k'}$	Somme
C_1	m_{11}	m_{12}	...	$m_{1j'}$	X_1
C_2	m_{21}	m_{22}	...	$m_{2j'}$	X_2
...	...	...	...	...	...
C_k	$m_{j'1}$	$m_{j'2}$	...	$m_{kk'}$	X_k
Somme	Y_1	Y_2	...	$Y_{k'}$	

Table 2.1 – Matrice de confusion illustrant l'information mutuelle entre le partitionnement automatique (P) et la classification experte (φ).

Cette indice varie dans un intervalle de $[0, 1]$, plus élevée est la valeur de cette indice meilleure est la qualité du partitionnement généré.

— L'Indice de Rand Ajusté : L'Indice de Rand Ajusté (ARI) mesure la similarité entre deux partitionnements en considérant le regroupement de toutes les paires d'entités (E_i *et* $E_{i'}$) de l'ensemble $\mathbb{E}$. Il identifie les paires d'entités regroupées ensemble ainsi que les paires d'entités regroupées séparément dans les deux partitionnements. Le ARI est défini par l'équation 2.7.

$$ARI(P, \varphi) = \frac{a - Exp[a]}{max(a) - Exp[a]} \tag{2.7}$$

Où :

— a représente le nombre de paires d'entités regroupées ensemble dans les deux partitionnements.

— $Exp[a]$ représente la valeur probable de a.

— $max(a)$ représente la valeur maximale de a.

Les valeurs $E[a]$ et $max(a)$ sont définis par les Équations 2.8 et 2.9.

$$Exp[a] = \frac{\pi(P) \cdot \pi(\varphi)}{n(n-1)/2} \tag{2.8}$$

$$max(a) = \frac{1}{2}\left(\pi(P) + \pi(\varphi)\right) \tag{2.9}$$

Où :

— $\pi(P)$ représente le nombre de paires d'entités regroupées dans le même cluster dans P.

— $\pi(\varphi)$ représente le nombre de paires d'entités regroupées dans la même classe dans φ.

L'indice de rand ajusté varie dans un intervalle $[0, 1]$. Lorsque les deux partitionnements concordent parfaitement, l'ARI est égal à 1.

2.3 Les images satellites

Les images satellites sont des images matricielles qui décrivent la surface de la terre ou une partie de celle-ci, prise par un capteur embarqué sur un satellite en orbite. Elles font partie de la famille des images d'observation de la terre comprenant les photos aériennes, les images radar etc.

Les images satellites sont composées d'un ensemble d'unités de base appelées pixels. Chaque pixel est caractérisé par une ou plusieurs valeurs. Cette valeur correspond à l'intensité de la lumière reflétée par la surface de la Terre.

La Figure 2.3 illustre un exemple d'image satellite représentée sous forme matricielle. Chaque case de la matrice représente un pixel et chaque pixel est caractérisé par une valeur radiométrique.

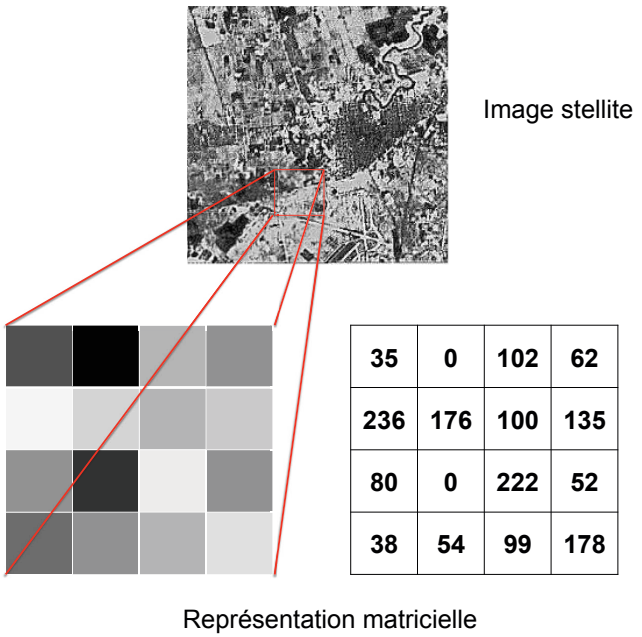


FIGURE 2.3 – Représentation matricielle d'une image satellite.

La taille de la superficie que le pixel représente correspond à la résolution spatiale de l'image. On distingue les images à très haute résolution ayant une résolution de 5 mètres ou moins, les images à haute résolution ayant une résolution de 30 à 10 mètres, les image à résolution moyenne ayant une résolution de 80 mètres et les image à base résolution ayant une résolution de 1000 mètres.

2.3.1 Acquisition des images satellites

La télédétection est une discipline qui permet l'acquisition des images satellites. Elle est définie comme l'ensemble des connaissances et techniques utilisées pour déterminer des caractéristiques physiques et biologiques d'objets par des mesures effectuées à distance, sans contact matériel avec ces objets. La télédétection appliquée à l'observation de la Terre implique l'usage de satellite artificiel et de capteur, afin de capter le rayonnement de la surface de la terre. On distingue deux types de capteurs : les capteurs actifs et capteurs passifs. Les capteurs passifs exploitent une source de lumière naturelle principalement le rayonnement solaire tandis que les capteurs actifs sont dotés de leur propre source de lumière.

Nous traitons d'images satellites optiques issues de capteurs passifs, le processus d'acquisition de ces images compte six étapes. Tout d'abord le soleil (source d'énergie) envoie des rayonnements en direction de la surface de la terre (cible) à observer (étape 1). Ces rayonnements traversent l'atmosphère pour atteindre la cible (étape 2). La cible interagit avec les rayonnements, elle absorbe une partie et réfléchit une autre partie selon ces propriétés (étape 3). Les rayonnements réfléchis sont captés

à distance pour être enfin enregistrés par le capteur (étape 4). Le capteur transmet ces rayonnements à une station de réception (étape 5). La station de réception transforme cette information en images satellites (étape 6)

2.3.2 Rayonnement électromagnétique

En télédétection, on exploite les propriétés du rayonnement. Le rayonnement est défini par des ondes électromagnétiques. Ces ondes sont caractérisées par une fréquence (Hz) ou une longueur d'onde (mètre). La Figure 2.4 illustre le spectre électromagnétique décrivant les différentes longueurs d'ondes du rayonnement.

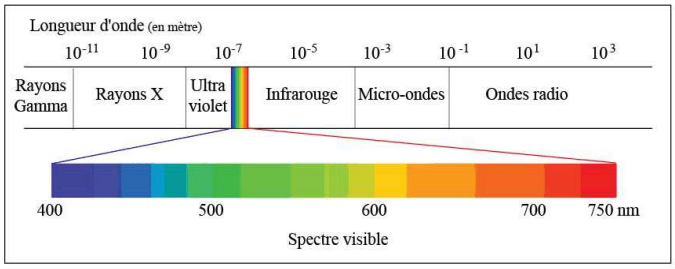


FIGURE 2.4 – Spectre électromagnétique.

Le spectre est divisé en fenêtres spectrales appelées bandes spectrales. Nous distinguons le spectre visible qui représente les rayonnements perçus par l'œil humain, l'infrarouge, l'ultraviolet etc. La surface de la Terre réfléchit les rayonnements avec différentes intensités selon sa couverture. La Figure 2.5 permet d'observer que la végétation absorbe les rayonnements visibles et réfléchit les rayonnements proche infrarouge, de même pour le sol nu. Tandis que les surfaces en eau réfléchissent les rayonnements visibles.

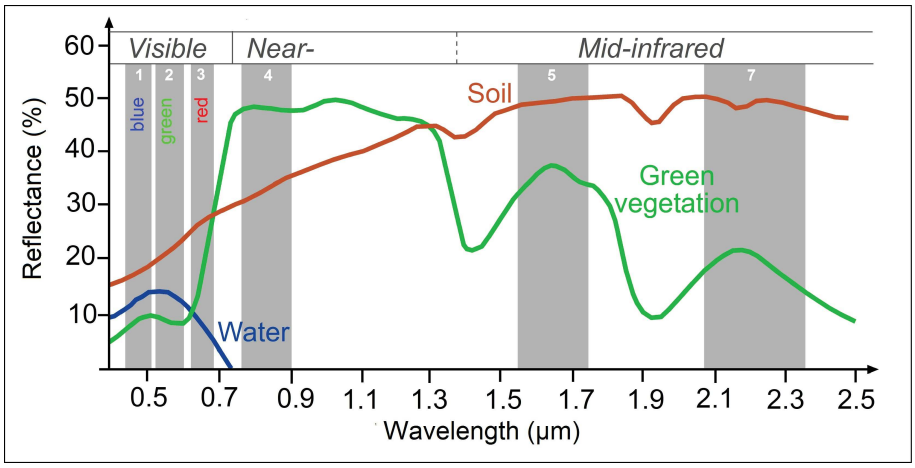


FIGURE 2.5 – Signatures spectrale de l'eau, de la végétation verte et du sol sur différentes fenêtres du spectres électromagnétiques.

Les satellites sont caractérisés par le nombre de bandes spectrales qu'ils sont capables d'observer et les intervalles de longueurs d'ondes de chacune de ses bandes. Le nombre de bandes spectrales définit la résolution spectrale de l'image satellite produite. Nous distinguons les images monospectrales, multispectrales et hyperspectrales.

Les images monospectrales sont définies par une seule bande spectrale, leurs pixels sont caractérisés par une seule valeur. Les images multispectrales sont définies par plusieurs bandes spectrales (3 à 10 bandes spectrales) tandis que les images hyperspectrales sont définies par plus d'une dizaine de bandes spectrales (plus de 10 bandes spectrales). Leurs pixels sont caractérisés par un vecteur de valeurs radiométriques.

La Figure 2.6 illustre la représentation d'une image multispectrale.

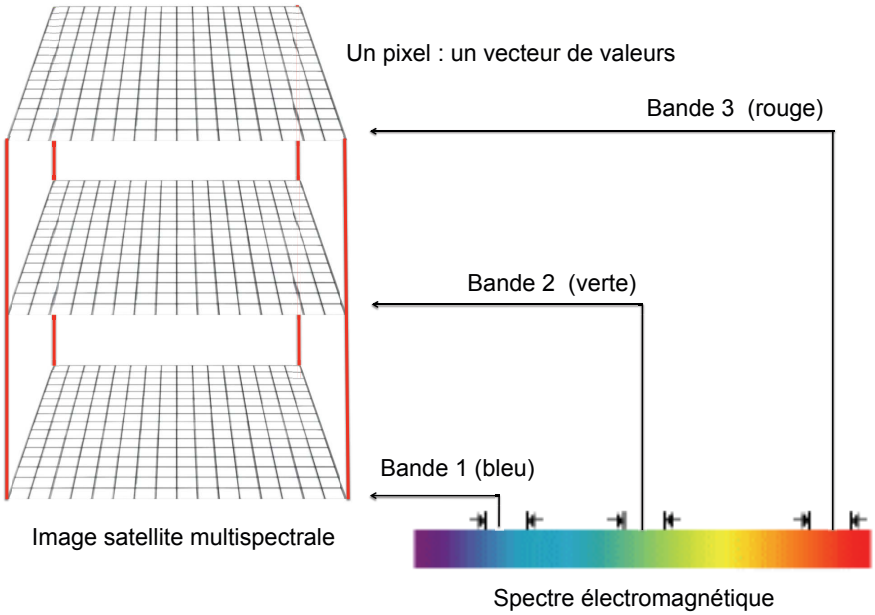


FIGURE 2.6 – Représentation d'une image multispectrale.

2.3.3 Satellite d'observation de la terre

Il existe de nos jours un nombre important de satellites artificiels en orbite permettant d'acquérir de grandes quantités d'images satellites. Certaines de ces images satellites sont mises gratuitement à disposition des chercheurs. Dans nos travaux de thèse, nous avons exploité des images Spot (2,4 et 5) et des images Landsat-5. Chacun de ces satellites possède des caractéristiques spécifiques à leur capteur. Nous allons vous présenter les caractéristiques de ces deux satellites. Nous décrirons aussi les satellites Sentinel-2 étant donné qu'il représente une famille de satellites récente offrant des images satellites à haute résolution spatiale.

— Image Landsat : Landsat est un programme spatial d'observation de la terre initié par le gouvernement des États-Unis et la NASA, destiné à des fins civiles. Il a commencé avec le lancement du premier satellite Landsat-1 en 1972. Sept autres satellites Landsat ont été lancés par la suite entre 1972 et 2013 : Landsat-2, Landsat-3, Landsat-4, Landsat-5, Landsat-6, Landsat-7 et Landsat-8.

La figure 2.7 illustre les satellites Landsat et la durée de chacune de leur mission :

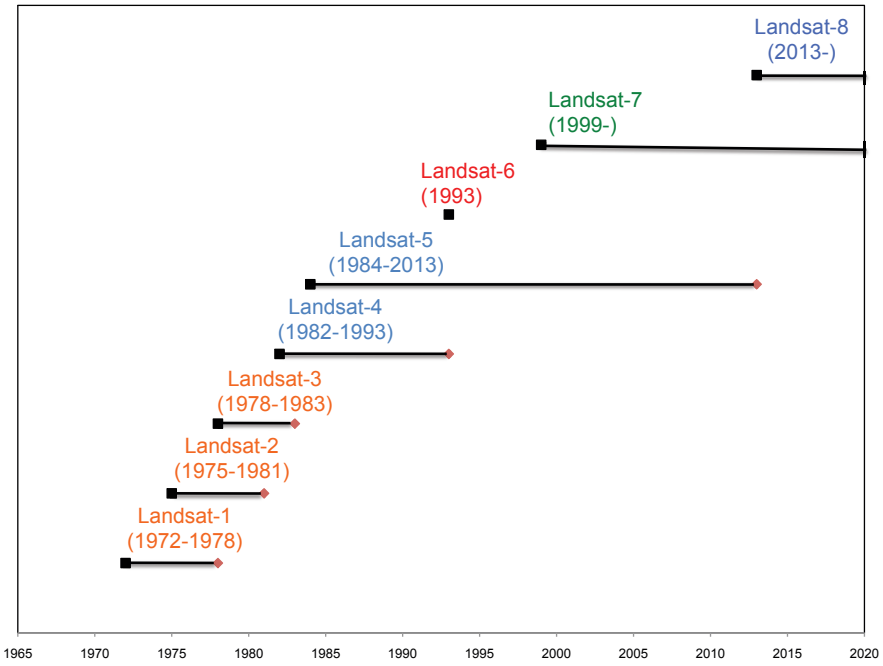


FIGURE 2.7 – Les satellites du programme Landsat.

Les images Landsat sont des images à haute résolution. Les caractéristiques des images Landsat-5 sont reportées dans le tableau 2.2 :

— Image Spot : Le Système Probatoire d'Observation de la Terre, dit Spot, est un programme civil pour l'observation de la terre. Il est développé par le centre national d'étude spatiales (CNES) français. Spot compte une famille de sept satellites : Spot-1, Spot-2, Spot-3, Spot-4, Spot-5, Spot-6, et Spot-7 (IMAGE, 1988). La chronologie des missions des satellites Spot est illustrée dans la figure 2.8.

Les image Spot sont des images à haute résolution. Le Tableau 2.3 décrit les images Spot-2, Spot-4, et Spot 5.

— Image Sentinel : Sentinel est une famille de satellites d'observation de la terre. Ils font partie du programme Copernicus de l'Union européenne en collaboration avec l'agence spatiale européenne. Ils sont conçus dans l'objectif de mettre

TABLE 2.2 – Principales caractéristiques des satellites Landsat-5.

Caractéristiques	Landsat-5
Début et fin de mission	1984-2013
Résolution	30 m en multispectral
Bandes spectrales	Bleu (0,45-0,52 μm) Vert (0,52-0,6 μm) Rouge (0,63-0,69 μm) Proche infrarouge (0,76-0,9 μm) Moyen infrarouge 1 (1,55-1,75 μm) Moyen infrarouge 2 (2,08-2,35 μm)

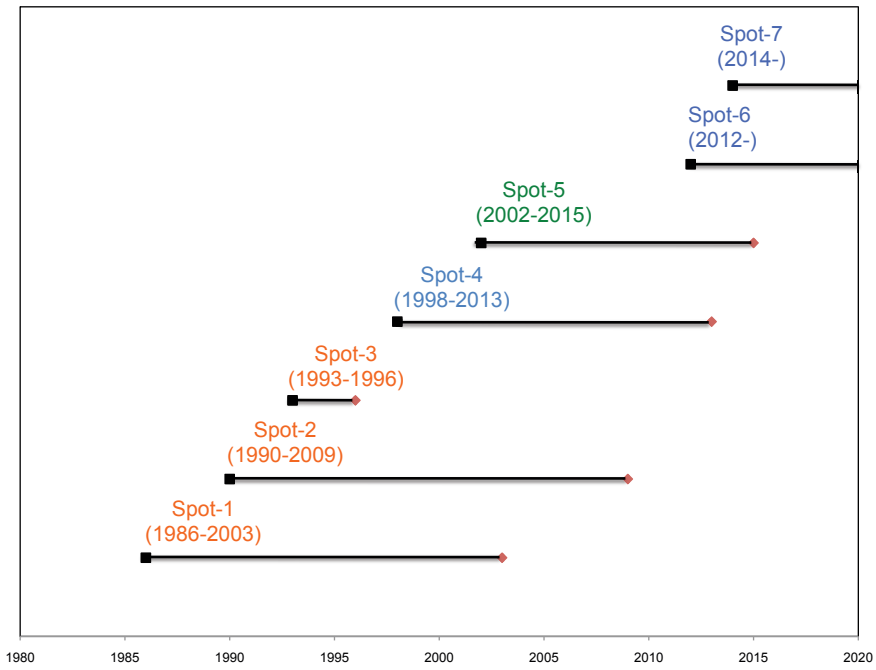


FIGURE 2.8 – Les satellites du programme Spot.

à disposition des pays européens de manière normalisée et continue des informations sur le sol, les océans et l'atmosphère du globe terrestre (SENTINEL, 2013). Ces informations permettent d'effectuer une multitude de taches telles que : la surveillance de l'environnement, la gestion des risques et la cartographie du sol. Les satellites Sentinel sont composés de cinq groupes comprenant les satellites : Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-4 et Sentinel-5. La chronologie des missions des premiers satellites Sentinel de chaque groupe est illustrée dans la Figure 2.9.

Ces satellites fournissent différentes informations spatiales dont des images ra-

TABLE 2.3 – Principales caractéristiques des satellites Spot-2, Spot-4, et Spot-5.

Caractéristiques	Sopt-2	Spot-4	Spot-5
Début et fin de mission	1990-2009	1998-2013	2002-2015
Résolution	20 m en multispectral	20 m en multispectral	10 en multispectral
Bandes spectrales	Vert (0,50-0,59 μm) Rouge (0,61-0,68 μm) Proche infrarouge (0,78-0,89 μm)	Vert (0,50-0,59 μm) Rouge (0,61-0,68 μm) Proche infrarouge (0,78-0,89 μm) Moyen Infrarouge (1,58-1,75 μm)	Vert (0,50-0,59 μm) Rouge (0,61-0,68 μm) Proche infrarouge (0,78-0,89 μm) Moyen Infrarouge (1,58-1,75 μm) à 20 m

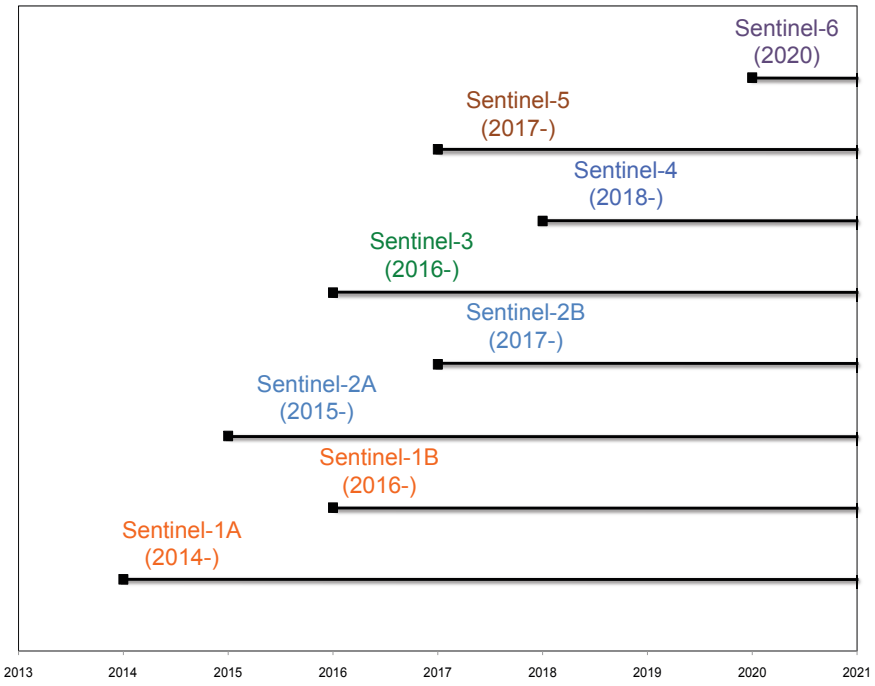


FIGURE 2.9 – Les satellites du programme Sentinel.

dar fournies par Sentinel-1 et des images mulispectrales fournies par Sentinel-2. On compte deux satellites Sentinel-2 : Sentinel-2A et Sentinel-2B. Ils fournissent des images spectrales à haute résolution spatial (10 m). La Table 2.4 décrit les caractéristiques des satellites Sentinel-2

Les satellites permettent d'acquérir plusieurs images d'une même zone à des dates différentes. Ces images constituent une série temporelle. Les séries temporelles d'images satellites (STIS) sont caractérisées par une résolution temporelle. La résolution temporelle est définie par la période de revisite d'une même zone par le satellite. Les satellites Sentinel sont caractérisés par une haute résolution temporelle

TABLE 2.4 – Principales caractéristiques des satellites Sentinel-2 (2A et 2B).

Caractéristiques	Sentinel-2A	Sentinel-2B
Début et fin de mission	2015-	2017-
Résolution	10 m	10 m
Bandes spectrales	Bleu (0.447 μm- 0.545 μm)	Bleu (0.443 μm- 0.541 μm)
	Vert (0.537 μm- 0.582 μm)	Vert (0.536 μm- 0.582 μm)
	Rouge (0.645 μm- 0.683 μm)	Rouge (0.645 μm- 0.684 μm)
	Red edge 1 (0.694 μm- 0.713 μm) (20 m)	Red edge 1 (0.693 μm- 0.713 μm) (20 m)
	Red edge 2 (0.731 μm- 0.749 μm) (20 m)	Red edge 2 (0.73 μm- 0.748 μm) (20 m)
	Red edge 3 (0.768 μm- 0.796 μm) (20 m)	Red edge 3 (0.765 μm- 0.793 μm) (20 m)
	Proche infrarouge (0.762 μm- 0.907 μm)	Proche infrarouge (0.806 μm- 0.899 μm)
	Red edge 4 (0.848 μm- 0.881 μm) (20 m)	Red edge 4 (0.848 μm- 0.88 μm) (20 m)
	Moyen infrarouge 1 (1.542 μm- 1.685 μm) (20 m)	Moyen infrarouge 1 (1.539 μm- 1.68 μm) (20 m)
	Moyen infrarouge 2 (2.081 μm- 2.323 μm)	Moyen Infrarouge 2 (2.066 μm- 2.304 μm) (20 m)

où la période de revisite est de cinq jours.

Dans cette section, nous avons défini les concepts de base en télédétection et traitement d'image satellite, ces concepts sont détaillées dans l'ouvrage de (GIRARD et GIRARD, 2010)

2.4 Analyse des séries temporelles d'images satellites

Les méthodes d'apprentissage automatique ont été largement utilisées dans le domaine de l'analyse des images satellites incluant les méthodes supervisées (JIANG et collab., 2013), les méthodes non supervisées (KURTZ et collab., 2010 ; WEMMERT et collab., 2009) ainsi que les méthode semi-supervisées (CHAHDI et collab., 2016 ; TAN et collab., 2016). Les méthodes d'apprentissage supervisé requièrent un ensemble de données étiquetées, l'étiquetage des images satellites est une tache qui requière un effort humain et un temps de traitement considérable. Afin de palier cette problématique, des méthodes d'apprentissage non supervisé ont été exploitées dans le domaine de l'analyse des images satellites.

Le clustering des images satellites repose sur deux facteurs qui sont : (i) le choix de la mesure de distance et (ii) la représentation des données. Concernant le choix de la mesure de distance, les travaux de la littérature utilisent généralement la distance euclidienne ou une de ses variantes (DING et collab., 2008). L'utilisation de cette mesure est favorisée car elle est intuitive, facile à implémenter et sans paramètre. Cependant le choix de la représentation des images satellites au niveau pixel ou bien au niveau objet suscite l'intérêt de la communauté scientifique (GUTTLER et collab.,

2017 ; Petitjean et collab., 2012b). La quelle des deux représentations est la plus adaptée au traitement d'images satellites permettant de les caractériser et d'exploiter leur information ? Nous présentons dans ce qui suit les différents travaux d'analyse d'images satellites basés pixel et ceux basés objet.

2.4.1 Analyse basée pixel

Les méthodes d'analyse par apprentissage automatique basées pixel traitent des images satellites au niveau du pixel. Les pixels représentent l'unité de base permettant de définir une image satellite. L'image se présente ainsi sous forme d'un ensemble de pixels localisés spatialement.

Ces méthodes identifient d'abord les pixels dans des images satellites puis les caractérisent en utilisant des descripteurs radiométriques. On distingue deux types d'information radiométriques : Les bandes spectrales et les indices radiométriques. Les bandes spectrales représentent l'information brute portée par le pixel. Les indices radiométriques sont des descripteurs calculés à partir de la combinaison de plusieurs bandes spectrales. Ils permettent de caractériser les différents types d'occupation du sol tels que l'indice de végétation qui permet de caractériser la végétation et le l'indice d'eau qui permet de caractériser les surfaces on eau. Enfin, ces descripteurs sont exploités dans le processus d'apprentissage supervisé/non supervisé afin de regrouper les pixels similaires comme illustré dans la Figure 2.10.

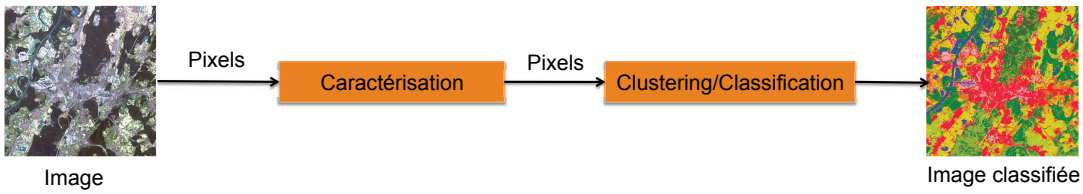


Figure 2.10 – Processus d'analyse d'image satellite basée pixel (Chahdi, 2017).

Nous nous intéressons à l'analyse de séries temporelles d'images satellites. Ce type de données est caractérisé non seulement par une dimension spatiale mais aussi par une dimension temporelle. L'analyse basée pixel des séries temporelles d'images satellites repose sur le caractérisation des pixels par rapport à tous les images de la série. Ainsi pour chaque pixel, on identifie ses valeurs radiométriques sur chacune des images de la série temporelle. Ces valeurs radiométriques du pixel sont regroupées et ordonnées en se basant sur la dimension temporelle des images satellites. Nous obtenons pour chaque pixel une série temporelle de valeurs radiométriques (Petitjean et collab., 2012b).

Dans la littérature, les séries temporelles sont exploitées dans différentes tâches. On en distingue principalement la cartographie du sol (Gonçalves et collab., 2014 ; Petitjean et collab., 2012a; 2011 ; Schuster et collab., 2015 ; Zhang et collab.,

2016), la détection de changements et l'évolution de la surface du sol (Guyet et Nicolas, 2016 ; Romani et collab., 2011).

Ces approches permettent de résumer les connaissances contenues dans les séries temporelles dans une seule image, dans le cas de la cartographie du sol, on obtient une carte d'occupation du sol, dans le cas de la détection de changements, on obtient une carte de changements (Julea et collab., 2011).

Dans les travaux de (Zhang et collab., 2016), les auteurs proposent une méthode d'analyse d'images satellites pour la cartographie du sol. Ils calculent une matrice de distance en utilisant la distance euclidienne entre les attributs des pixels. Cette matrice est ensuite exploité par l'algorithme de clustering de propagation d'affinité (Frey et Dueck, 2007) afin d'identifier les différentes classes d'occupations du sol. Les performances du clustering de propagation d'affinité ont été comparées aux performances du Kmeans et de l'algorithme hiérarchique agglomératif sur trois zones d'étude différentes. Les résultats ont montré que l'algorithme de propagation d'affinité obtient de meilleures performances. À la différence de la méthode précédente, (Petitjean et collab., 2012a; 2011) introduit une méthode de cartographie basée sur la mesure Dynamic Time Warping (DTW)(Sakoe et Chiba, 1978). Cette méthode permet de gérer les séries à échantillonnage irrégulier et de tailles différentes pouvant contenir des valeurs manquantes (Image satellite contenant des nuages). En effet, DTW permet de palier cette problématique et d'aligner ces séries temporelles. Une fois la distance estimée entre les séries temporelles, ils sont regroupés en utilisant l'algorithme Kmeans. Les résultats expérimentaux démontrent que la mesure DTW est adéquate pour l'estimation des distances entre les séries temporelles d'images satellites et permet de générer un regroupement cohérent. De même les auteurs (Gonçalves et collab., 2014) proposent dans leurs travaux une méthode de cartographie du sol basée sur la mesure DTW et l'algorithme de clustering Kmeans. (Schuster et collab., 2015) proposent une approche supervisée pour la cartographie des couverts végétaux dans les zones semi-naturelles en utilisant des séries temporelles d'indice de végétation. Ces séries sont ensuite classées en utilisant les machines à vecteurs de support (Schölkopf et collab., 2002).

Pour le suivi de l'évolution de la couverture du sol, les travaux de (Guyet et Nicolas, 2016) proposent une méthode d'analyse de STIS qui permet d'observer les couverts végétaux. Cette méthode exploite des STIS pluriannuelles, chaque pixel de la série est caractérisé en utilisant l'indice de végétation. L'indice de végétation permet de construire une série temporelle annuelle. Ces séries sont analysées en utilisant l'algorithme de clustering Kmeans afin d'identifier des séries annuelles types. Les séries annuelles types de chaque pixel sont ensuite rassemblées afin de décrire l'évolution pluriannuelle du pixel. Une autre étape de clustering est réalisée sur les profils pluriannuels en utilisant l'algorithme Kmédoïdes. Afin d'estimer la distance entre ces séries, cette méthode utilise la distance euclidienne. Les résultats expérimentaux ont démontré que le clustering en deux étapes permet d'obtenir de meilleures performances. Dans (Romani et collab., 2011), les auteurs proposent une approche similaire qui se base sur l'analyse de séries temporelle afin de surveiller les

culture de canne à sucre. Ils exploitent l'indice de végétation pour caractériser les pixels. Puis, ils estiment la distance entre les séries temporelles construites en utilisant la mesure DTW. Ensuite, ils regroupent celles présentant la même évolution en utilisant l'algorithme Kmeans et Kmédoïdes. Cette approche permet d'identifier dans la zone d'étude les régions caractérisées par une forte production de canne à sucre ainsi que l'expansions de la canne à sucre vers différentes régions.

2.4.2 Analyse basée objet

Les méthodes d'analyse par apprentissage automatique basées objet permettent de traiter des unités structurelles des images telles que les parcelles agricoles, le bâtis, les rivières etc. Ces unités sont identifiées en utilisant des algorithmes de segmentation. Les algorithmes de segmentation (Dey et collab., 2010) analysent les pixels afin de les regrouper selon des critères d'homogénéité spectrale, texturale et spatiale. Les différents groupes de pixels forment des objets correspondant aux unités décrites par l'image satellite. Ces objets sont caractérisés en utilisant les valeurs radiométriques de leur pixels (moyenne des bandes spectrales et indices radiométriques) ainsi que des descripteurs géométriques (taille de l'objet etc). L'analyse basée objet consiste à réaliser le clustering sur les objets afin d'identifier ceux qui sont similaires. La Figure 2.11 illustre le processus d'analyse d'images satellites basée objet.

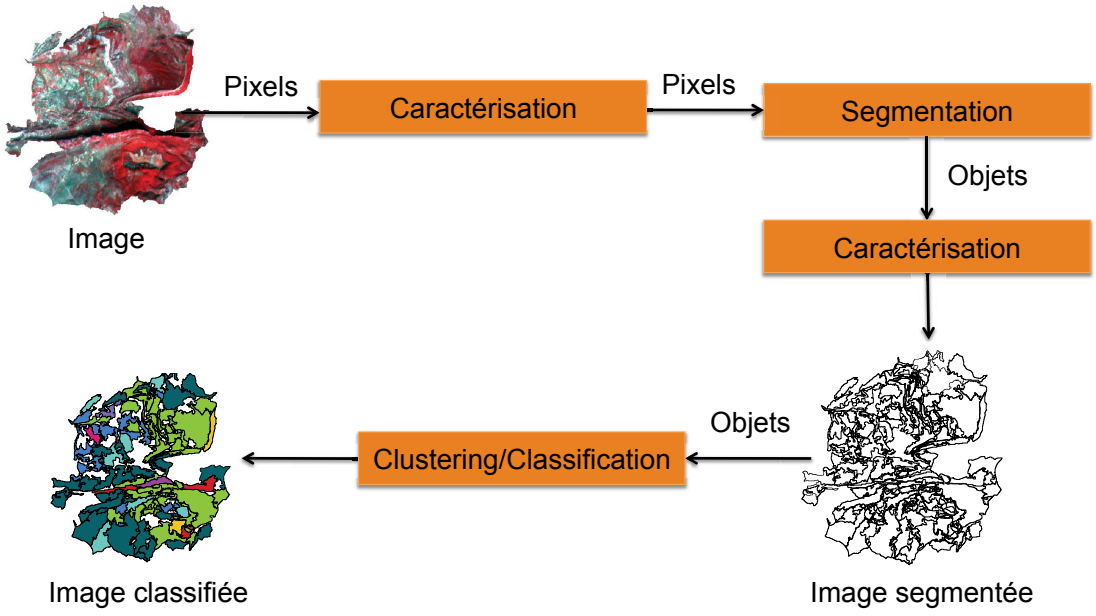


Figure 2.11 – Processus d'analyse d'image satellite basée objet (Chahdi, 2017).

L'analyse d'images satellites basée objet est connue sous le nom de OBIA (Object-Based Image Analysis) (Blaschke, 2010). Il existe dans la littérature un

grand nombre de travaux exploitant les méthodes basé objet pour l'analyse d'une seule image (zho ; Laliberte et collab., 2004 ; Mathieu et collab., 2007). Tandis qu'il existe peu de travaux traitant de l'analyse de séries temporelles d'images satellites. En effet, l'alignement de pixels dans les séries temporelles est une tache facile car il existe une correspondance entre les pixels des images satellites. Cependant l'alignement d'objets est une tâche non triviale car il n'existe pas de correspondance entre les objets des images de la série. En effet, le changement d'occupation du sol décrit dans les images satellites mène à la distorsion et à l'expansion des différents objets identifiés d'une image à une autre.

Parmi les premiers travaux proposés dans la littérature, les auteurs (Petitjean et collab., 2012b) introduisent une méthode combinant l'analyse basée pixel et l'analyse basée objet. Cette méthode considère d'abord les pixels de la série temporelle et les caractérise en utilisant leurs valeurs radiométriques. Elle segmente ensuite les images de la série temporelle afin d'identifier des objets qui sont caractérisés par leur information radiométrique et géométrique (taille de l'objet etc). Afin de combiner les deux représentations de l'image, les attributs du pixel sont enrichis avec les attributs de leur objet. Les séries temporelles résultats sont ensuite analysées en utilisant l'algorithme Kmeans, tandis que l'estimation des distances a été réalisée en utilisant la distance euclidienne. Les résultats ont démontré la pertinence de l'information contextuelle des pixels pour la tâche de cartographie du sol.

L'un des premiers travaux à proposer l'analyse de séries temporelles basée objet est introduit par (Desclée et collab., 2006). Les auteurs combinent les différentes images satellites afin d'effectuer une segmentation multi-dates. Les images satellites sont analysées afin d'identifier pour chaque pixel une série temporelle de valeurs radiométriques en combinant ses valeurs sur chaque image. Ces séries temporelles sont ensuite utilisées pour réaliser la segmentation. L'algorithme de segmentation regroupe les pixels homogènes. Cette méthode intègre les trois dimensions spatiale, spectrale et temporelle dans le processus de segmentation, afin de générer un même ensemble d'objets dans toutes les images satellites de la série. Afin d'identifier le changement de couvert forestier dans la série, la méthode considère les images satellites successives deux à deux. Les objets résultats de la segmentation sont caractérisés par rapport à ces deux images puis classés en utilisant des méthodes statistiques. De même dans leurs travaux (Qin et collab., 2013), les auteurs combinent deux images en empilant leurs bandes spectrales. L'image résultat est ensuite segmentée afin d'identifier les objets la constituant. Ces objets sont analysés à l'aide d'algorithmes de classification afin d'identifier les changements de l'occupation du sol. Des travaux similaires proposant des méthodes d'analyse basée objet en utilisant le segmentation muti-date ont été proposées par (Bontemps et collab., 2008 ; Conchedda et collab., 2008).

Récemment, (Guttler et collab., 2017) propose une approche permettant d'observer les phénomènes spatio-temporels et d'extraire des profils d'évolution à partir des STIS. Cette méthode vise à détecter et à extraire automatiquement des profils spatio-temporels des STIS. Ils segmentent chaque image satellite indépendamment,

puis relie les objets identifiés sur chaque deux images successives de la série temporelle. La méthode permet de construire un ensemble de profils temporels décrivant les l'évolutions spatiale et radiométrique des objets.

2.5 Conclusion

Dans ce chapitre, nous avons introduit les concepts nécessaires pour la compréhension de nos travaux de thèse. Nous avons d'abord défini les différentes méthodes d'apprentissage automatique, nous avons présenté quelques concepts de base en télédétection et enfin nous avons décrit les différents travaux d'analyse des images satellites par apprentissage automatique.

Nous avons défini trois méthodes d'apprentissage automatique : supervisé, non supervisé et semi-supervisé. L'apprentissage supervisé se base sur des données étiquetées pour construire des modèles permettant de classer les données. Les données étiquetées sont souvent non disponibles ou requièrent un effort humain considérable. Pour palier cette problématique, nous avons introduit les méthodes d'apprentissage non supervisé. Ces dernières n'utilisent pas de données étiquetées, elles permettent d'identifier des groupes cohérents de données. Nous avons présenté les différentes familles d'algorithmes non supervisés et nous avons détaillé ceux exploités durant cette thèse : le clustering hiérarchique agglomératif et le clustering spectral. Nous avons aussi défini les mesures de qualité internes et externes permettant d'évaluer les performances des ces algorithmes. Enfin nous avons introduit les méthodes semi-supervisés qui permet à la fois d'utiliser des données étiquetées et non étiquetées.

Nous avons défini les images satellites, leur processus d'acquisition et leurs caractéristiques. Les séries temporelles d'images satellites permettent d'observer l'évolution de la surface de la terre. Ces images peuvent être représentées par des pixels. Elles peuvent être aussi représentées par des objets, les objets correspondent à un ensemble de pixels homogènes. Nous avons présenté les différents travaux d'analyse par apprentissage automatique des séries temporelles basé sur ces deux représentations.

Dans le chapitre suivant nous décrirons nos zones d'étude, les images satellites exploitées pour leur analyse ainsi que leur pré-traitement.

Chapitre 3

Les données d'application

Sommaire

3.1 Introduction

Dans le Chapitre 2, nous avons défini les images satellites et décrit leur processus d'acquisition. Nous avons aussi présenté différents travaux de la littérature exploitant les images satellites pour la cartographie du sol, la détection de changement de couverture du sol, etc. En effet, l'imagerie satellite est devenue une source incontournable d'information. Elle permet de répondre à diverses problématiques dans différents domaines environnementaux et sociétaux dont : la santé humaine, l'agriculture, le climat, l'énergie, les incendies, les catastrophes naturelles, la croissance urbaine, la gestion de l'eau, les écosystèmes, la biodiversité, etc. L'acquisition des images satellites au travers des satellites d'observation de la terre a commencé depuis les années cinquante. Ce domaine a depuis connu une effervescence, plusieurs satellites ont été mis en orbite par différents états et organismes. Aujourd'hui nous comptant plus de 5500 satellites en orbite. Plusieurs de ces organismes permettent l'accès gratuit aux images satellites dont les images Landsat, Sentinel, etc.

Dans cette thèse, nous nous intéressons à l'analyse d'images satellites pour le suivi de l'évolution des habitats naturels. Nous exploitons les séries temporelles d'images satellites pour l'observation de l'évolution de la couverture du sol.

Dans ce chapitre, nous introduisons tout d'abord les sites d'étude (Section 2.2). Nous présentons par la suite les images satellites qui permettrons d'analyser l'évolution de ces sites (Section 2.3). Nous décrirons également l'étape de pré-traitement de ces images satellites i.e., le calcul d'indices spectraux et la segmentation (Section 2.4). Nous finissons par présenter les cartes d'occupation de ces sites, qui ont été réalisées par un expert du terrain (Section 2.5).

3.2 Sites d'étude

Dans le cadre de cette thèse, notre analyse a porté sur quatre sites d'étude : (A) la Basse Plaine de l'Aude (BPA) , (B) la Vallée du Libron (VL), (C) la Montagne de la Moure et Causse d'Aumelas (MMCA) et (D) le Pic Saint Loup (PSL). Les Figures 3.2 et 3.1 montrent la localisation de ces sites.

Les quatre sites sont localisés aux sud de la France (Figure 3.1). La Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas et le Pic Saint Loup font partie de la liste des sites Natura 2000 [1]. Natura 2000 est un réseau qui groupe plus de 2500 site naturels ou semi-naturels de l'union européenne dont le but est de préserver leur diversité biologique (faune et flore).

3.2.1 La Basse Plaine de l'Aude

La Basse Plaine de l'Aude s'étend sur 4500 hectares (Figure 3.3). Localisée au sud de la France, plus précisément sur le littoral languedocien entre le massif de la Clape

1. `https://inpn.mnhn.fr/site/natura2000/listeSites`

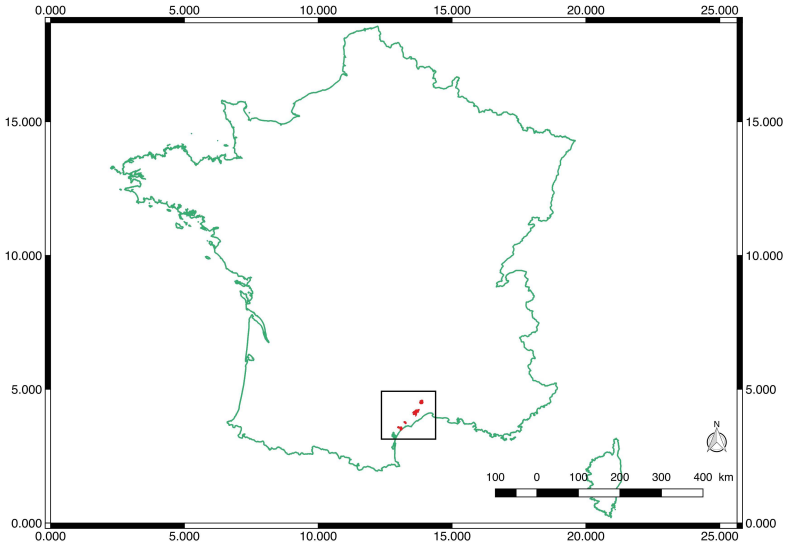


FIGURE 3.1 – Localisation des quatre sites d'étude au sud de la France.

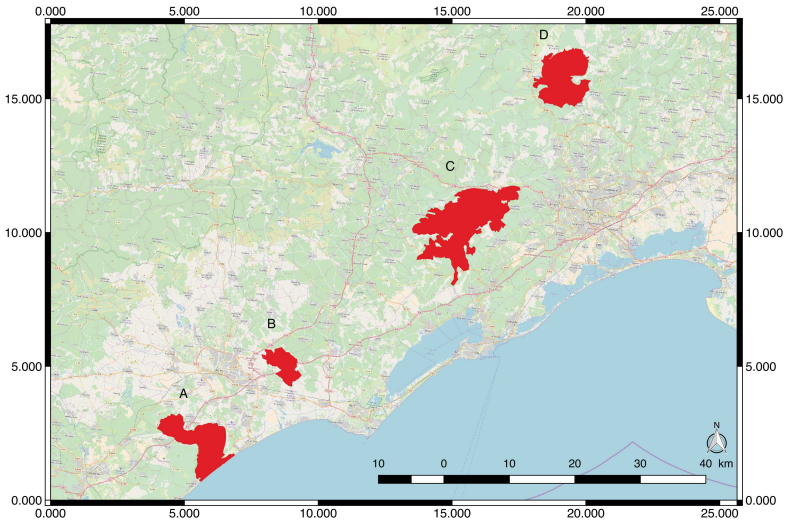


FIGURE 3.2 – Localisation des sites d'étude : (A) la Basse Plaine de l'Aude, (B) la Vallée du Libron, (C) la Montagne de la Moure et Causse d'Aumelas et (D) le Pic Saint Loup.

au sud et les collines du Narbonnais au nord. Elle est partagée par les départements de l'Aude au sud (avec les communes de l'agglomération du Narbonnais) et de l'Hérault au nord (avec les communes de l'agglomération de la Domitienne).

Traversée par la rivière de l'Aude, ce site est caractérisé par ses zones humides. En effet, il compte de vastes étangs : l'étang de Vendres, l'étang de la Matte et l'étang de Pissevaches qui couvrent plus de 40% de sa superficie. On y est trouve aussi des terres agricoles qui couvrent environ 50% de la superficie du ce site, elles

se composent principalement de vigne, de céréales ainsi que de prairies temporaires ou permanentes. Le site inclut également des dunes littorales constituées par la zone sableuse de l'étang de Pissevaches, l'embouchure de l'Aude et la plage de Vendres. Ce site compte ainsi une grande diversité de milieux naturels qui ont longtemps été menacés par des projets touristiques, aujourd'hui abandonnés. La Basse Plaine de l'Aude représente un lieu important pour les espèces nicheuses ainsi que les espèces migratrices qui le fréquentent soit pour s'y reproduire, soit pour hiverner ou comme lieu de halte migratoire.

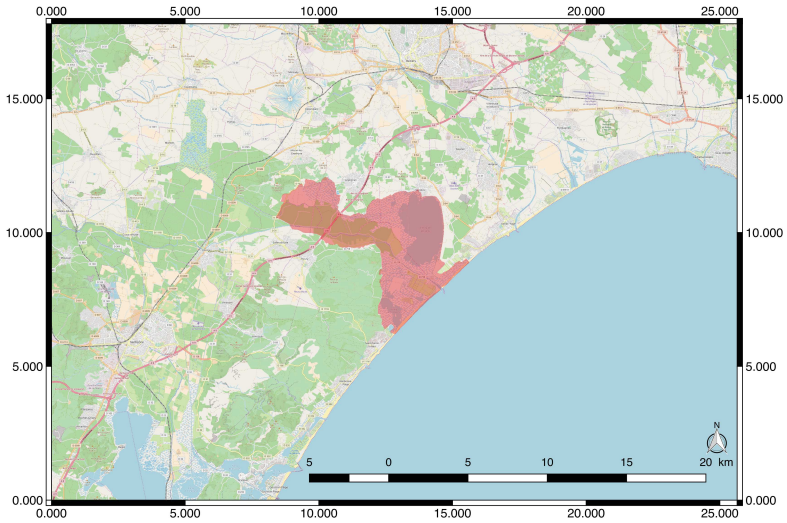


FIGURE 3.3 – Localisation de la Basse Plaine de l'Aude.

3.2.2 La Vallée du Libron

Localisée au sud de la France, à 10 kilomètres au nord-est de Bézier, la Vallée du Libron s'étend sur 1,655 hectares (Figure 3.4), elle est traversée par une petite rivière nommée le Libron.

Ce site est principalement composé de parcelles agricoles qui sont localisées près de la rivière Libron. On y trouve deux types de cultures prédominantes, les céréales au nord-ouest du site et les vignes au sud-est du site. La Vallée du Libron compte aussi des zones naturelles essentiellement composées de forêts et de garrigues. Elle comprend aussi un terrain de golf situé au nord. Dans ce site, la taille des terrains agricoles et les parcelles forestières est généralement de 6 à 8 ha (c.-à-d. 200 m x 400 m ou plus pour la plupart des cultures) ce qui facilite la distinction entre leurs limites.

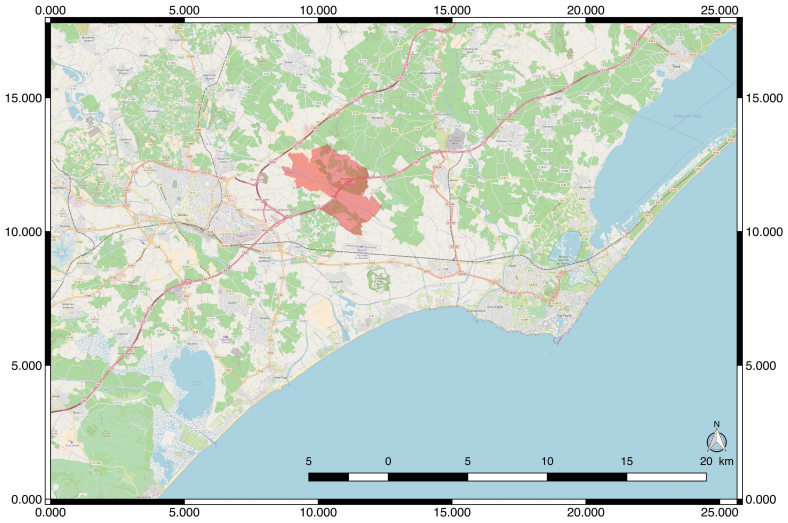


FIGURE 3.4 – Localisation de la Vallée du Libron.

3.2.3 La Montagne de la Moure et Causse d'Aumelas

La Montagne de la Moure et Causse d'Aumelas (MMCA) se situe au sud de la France, entre trois communautés d'agglomération : Montpellier Méditerranée à l'est, bassin de Thau au sud, vallée de l'Hérault au nord et à l'ouest. Ce site couvre une superficie de 9369 hectares (Figure 3.5).

La MMCA compte des habitats naturels et semi naturels. Elle est caractérisée par un paysage forestier. En effet la forêt couvre plus de 70% de sa superficie. Ces forêts sont principalement composées de garrigue, résineux, lande et broussailles. Ce site est aussi caractérisé par une activité pastorale ancienne. Cependant l'étalement urbain de l'agglomération de Montpellier et le risque d'abandon des pratiques pastorales traditionnelles se présentent comme deux facteurs menaçant la conservation de son équilibre naturel.

3.2.4 Le Pic Saint Loup

Localisé au sud de la France, le Pic Saint Loup se situe entre le massif des Cévennes au nord et l'agglomération de Montpellier au sud. D'une superficie de 4420 hectares (Figure 3.6), il compte 8 communes : Cazevieille, Mas de Londres, Notre Dame de Londres, Rouet, Saint Jean de Cuculles, Saint Martin de Londres, Saint Mathieu de Tréviers et Valflaunès.

Le Pic Saint Loup est l'un des points forts du paysage régional et présente une grande diversité d'habitats. Il compte des prairies, des forêts résineuses, des landes, des garrigues, et aussi des cultures. L'activité humaine telle que l'agriculture et le pastoralisme ont contribué au façonnement de son paysage. Ce site est marqué par les montagnes du Pic Saint-Loup (658 m) et de l'Hortus (512 m) qui rassemblent des

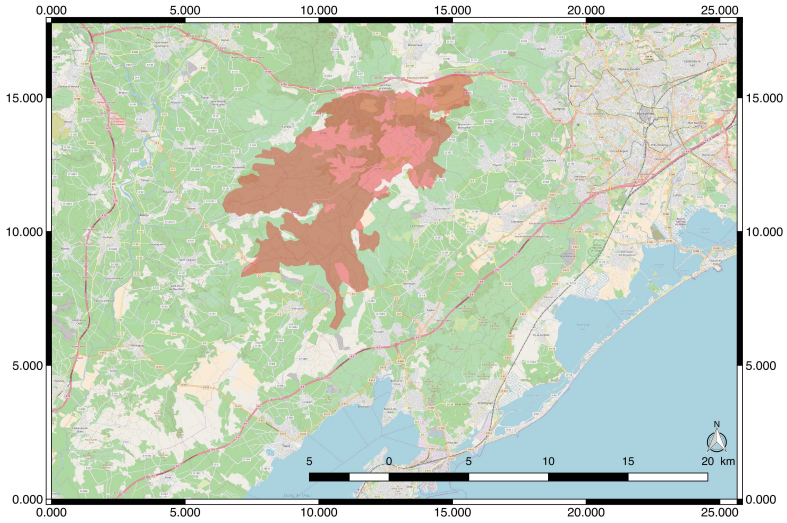


FIGURE 3.5 – Localisation de la Montagne de la Moure et Causse d'Aumelas.

espèces végétales d'affinité méditerranéenne à montagnarde et recèlent plusieurs espèces rares. De même que la MMCA, le Pic Saint Loup est menacé par l'abandon des pratiques pastorales et par une fréquentation croissante, en provenance notamment de l'agglomération montpelliéraine.

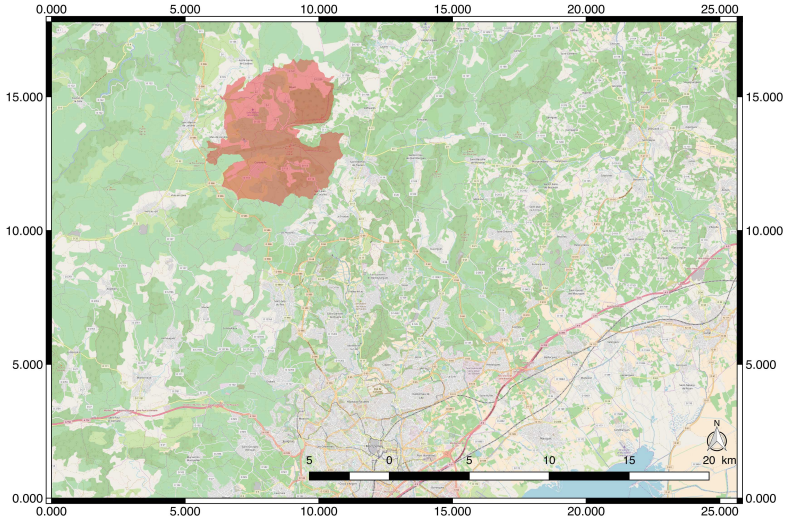


FIGURE 3.6 – Localisation du Pic Saint Loup.

3.3 Images satellites

Afin de pouvoir analyser et suivre l'évolution de nos quatre sites d'étude, nous avons constitué une base d'images satellites qui comporte 161 images issues de différents capteurs comprenant : Spot-2, Spot-4, Spot-5 et Landsat-5. Le Tableau 3.2 montre le nombre d'images par capteur acquises pour chacune de nos quatre zones d'étude.

	Spot-2	Spot-4	Spot-5	Landsat-5
BPA	29	16	8	6
MMCA	21	17	8	0
PSL	25	17	8	0
Vallée du Libron	0	0	0	6
Total	75	50	24	12

TABLE 3.1 – Le nombre d'images satellites acquises pour chaque zone d'étude et par capteur (Spot-2, Spot-4, Spot-5 et Landsat-5).

Les images Landsat-5 ont été acquises dans le cadre du programme Investissements d'Avenir du projet GEOSUD (http ://ids.equipex-geosud.fr/) supporté par l'agence nationale de la recherche. Tandis que les image Spot-2, Spot-4 et Spot-5 sont disponible dans le cadre du programme Spot World Heritage (SWH) soutenu par le CNES (Centre national d'études spatiales), de même les images Spot sont accessible via le site du pôle Théia : https ://theia-landsat.cnes.fr

Les images satellites Landsat et Spot possèdent des caractéristiques bien spécifiques à leur type de capteur (Section 2.3). Ces images satellites ont été organisées par capteur (Landsat et Spot) et par sites d'étude. Elles ont été réparties en cinq groupes, chaque groupe constitue une série temporelle d'images satellites : (i) Les images Landsat-5 ont été organisées en deux groupes de six images décrivant la Vallée du Libron et la Basse Plaine de l'Aude sur une période de huit mois. Ces deux séries temporelles ont permet une étude annuelle de l'évolution de la couverture de sol de la Vallée du Libron et la Basse Plaine de l'Aude sur une courte période. (ii) Les image Spot-2, Spot-4 et Spot-5 ont été reparties en trois groupes décrivant la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas, et le Pic Saint Loup sur une période de dix-huit ans. Ces séries temporelles ont été exploitées pour l'étude pluriannuelles de l'évolution de la couverture de sol de la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas, et le Pic Saint Loup.

3.3.1 Images satellites Landsat

Notre étude de l'évolution annuelle de la couverture du sol a été effectuée sur deux séries temporelles d'images satellites décrivant la Vallée du Libron et le Basse Plaine de l'Aude. Ces deux séries temporelles se composent de six images Landsat-5 acquises en 2009 qui couvrent une période de huit mois, de Février jusqu'en Septembre. Les

six bandes spectrales des images Landsat-5 ont été exploitées durant notre étude, à savoir les bandes : bleu, verte, rouge, proche infrarouge, moyen infrarouge 1, et moyen infrarouge 2. Les Figures 3.7 et 3.8 illustrent les images Landsat-5.

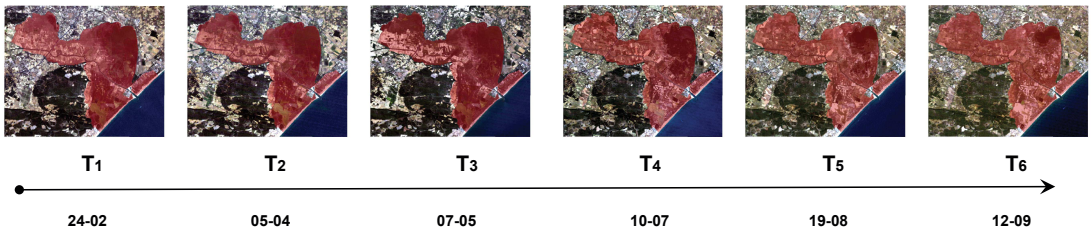


FIGURE 3.7 – Images satellites Landsat-5 acquises de la Basse Plaine de l'Aude.

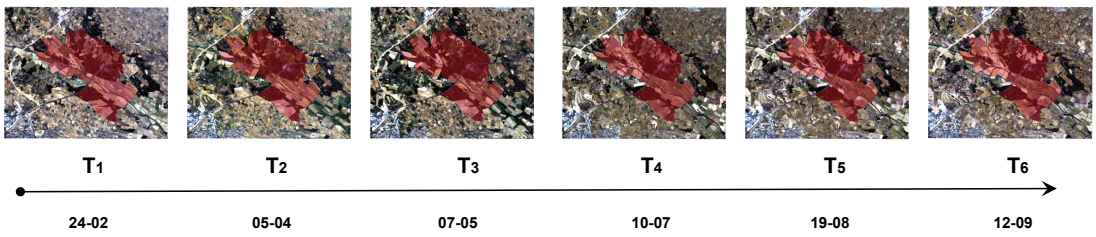


FIGURE 3.8 – Images satellites Landsat-5 acquises de la Vallée du Libron.

3.3.2 Images satellites Spot

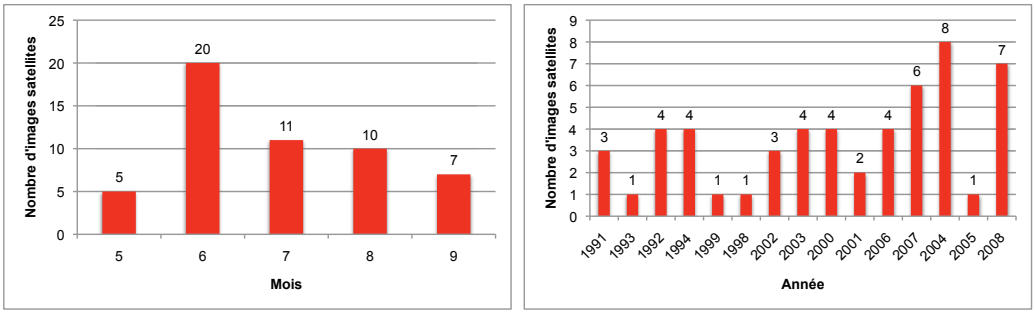
Notre base de données des séries temporelles d'images satellites pluriannuelles se compose de trois séries temporelles décrivant : la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas, et le Pic Saint Loup. Chacune des ces séries se compose de 53, 46, et 50 images respectivement. Les images décrivent les zones d'étude sur une période de dix-huit ans de 1990 à 2008.

Les Figure 3.9(b), 3.10(b), et 3.11(b) présentent la distribution des images satellites par année sur les trois sites : la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas, et le Pic Saint Loup respectivement. Quant aux Figure 3.9(a), 3.10(a), et 3.11(a), elles présentent la distribution des images satellite par mois pour chacun des trois sites. Cette période d'acquisition s'étend sur cinq mois, de Mai jusqu'en Septembre. Cette période offre des conditions atmosphériques adaptées pour l'acquisition d'images satellites car il y a moins de nuages durant ces cinq mois.

Les images satellites sont issues de trois capteurs différents dont Spot-2, Spot-4 et Spot-5 qui possèdent un nombre différents de bandes. Afin de constituer une base de données homogène, nous avons exploité les trois bandes spectrales : verte, rouge, et proche infrarouge communes aux quatre capteurs.

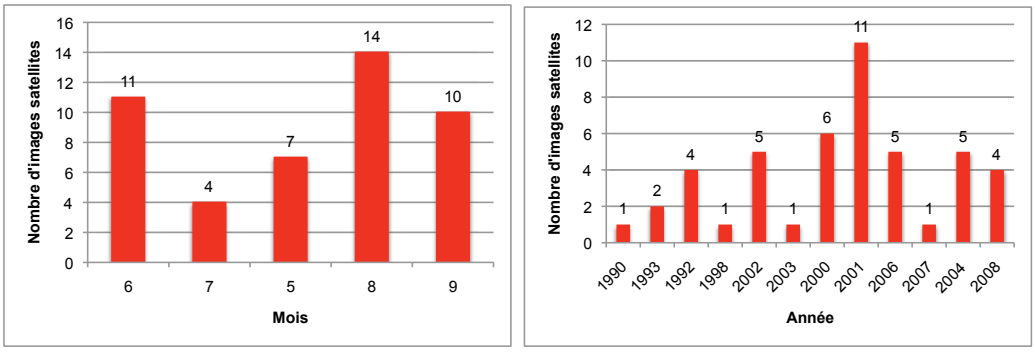
	Spot-2	Spot-4	Spot-5	Total
BPA	29	16	8	53
MMCA	21	17	8	46
PSL	25	17	8	50
Total	75	50	24	149

TABLE 3.2 – Le nombre d'images satellites acquises de : la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas, et le Pic Saint Loup par capteur : Spot-2, Spot-4 et Spot-5.



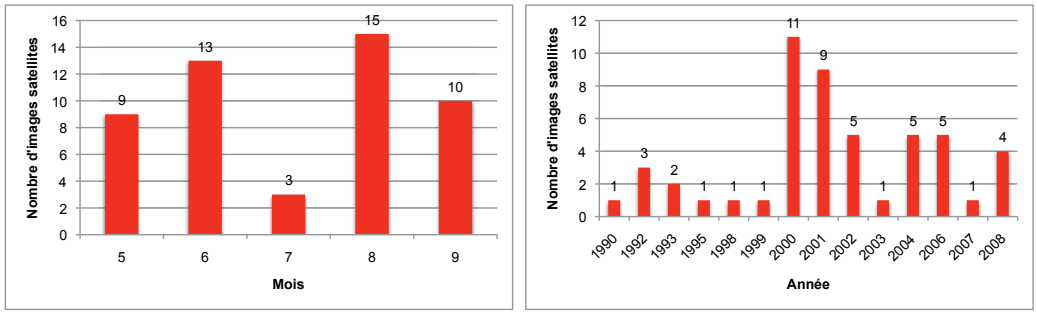
(a) (b)

FIGURE 3.9 – Nombre d'images satellites Spot2, Spot4 et Spot-5 acquises pour la Basse Plaine de l'Aude par année et par mois.



(a) (b)

FIGURE 3.10 – Nombre d'images satellites Spot2, Spot4 et Spot-5 acquises pour la Montagne de la Moure et Causse d'Aumelas par année et par mois.



(a) (b)

FIGURE 3.11 – Nombre d'images satellites Spot2, Spot4 et Spot-5 acquises du Pic Saint Loup par année et par mois.

3.4 Pré-traitement des images satellites

Le pré-traitement des images satellites consiste en deux étapes : le calcul d'indices spectraux et la segmentation. La première étape permet d'enrichir l'information spectrale des images satellites en calculant des indices spectraux. La deuxième étape permet d'identifier les unités structurelles des images satellites (parcelle agricole, forêt,...) et de leur générer une représentation objet.

3.4.1 Calcul d'indices spectraux

Notre base de données contient deux type de séries temporelles : les série temporelle Landsat et les séries temporelles Spot. Compte tenu des bandes spectrales caractéristiques de ces deux types de séries ainsi que la spécificité des sites d'étude qu'elles décrivent, nous avons calculé des indices spectraux différents. Les indices spectraux constituent une sur couche d'information spectrale et sont calculés à partir des bandes brutes. Ils exploitent les propriétés de réflectance des surfaces de sol afin de les caractériser.

3.4.1.1 Images satellites Landsat

Les série temporelles d'images satellites Landsat-5 sont décrit par six bandes spectrales : bleu, verte, rouge, proche Infrarouge, moyen infrarouge 1, et moyen infrarouge 2. L'information spectrale de ces images a été enrichie en calculant trois indices : indice de végétation par différence normalisé (NDVI), Indice de différence d'eau normalisé (NDWI) et indice de sécheresse visible et à ondes courtes (VSDI). Ces indices nous permettent de caractériser la végétation.

— Indice de végétation par différence normalisé (NDVI) (ROUSE JR et collab., 1974) : Le NDVI est calculé en considérant les deux bandes Rouge (R) et Proche Infrarouge (PIR) selon l'équation : $(PIR-R)$ / $(PIR+R)$. Cet indice

permet de caractériser le couvert végétal en exploitant ses propriétés qui sont la forte absorption dans le canal rouge et la forte réflectance dans le canal proche infrarouge. Les valeurs NDVI sont comprises dans l'intervalle [-1, 1]. Les valeurs négatives correspondent à des couvert non végétaux tels que l'eau et les nuages, quant à la valeur nulle, elle correspond au sol nu. Tandis que les valeurs positives correspondent à de la végétation, plus dense est le couvert végétal plus élevée est la valeur du NDVI.

— Indice de différence d'eau normalisé (NDWI) (Gao, 1996) : Le NDWI est estimé à partir des bandes Proche Infrarouge (PIR) et Moyen Infarrouge (MIR). Il est calculé de même que le NDVI : $(PIR - MIR)/(PIR + MIR)$. Il permet de surveiller la végétal dans les zones touchées par la sécheresse ainsi que la teneur en eau du couvert végétal. De même que le NDVI, il varie entre [-1,1]. Les valeurs négatives correspondent à un couvert végétal sec, quant aux valeurs positives, elles correspondent à un couvert végétal vert. En effet, le couvert végétal vert est caractérisé par une réflectance plus élevée dans le canal proche infrarouge par apport au canal moyen infrarouge.

— Indice de sécheresse visible et à ondes courtes (VSDI) (Zhang et collab., 2013) : Le VSDI est évalué en se basant sur les bandes Rouge (R), bleu (B), et Moyen Infrarouge (MIR). Il est défini par l'équation : $1 - (MIR - B) + (R - B)$. Cet indice est sensible au changement d'humidité, il permet de surveiller l'humidité du sol et de la végétation. En ce qui concerne la végétation, les bandes spectrales moyen infrarouge et rouge sont plus sensibles à la variation de l'humidité en comparaison à la bande bleu, ainsi calculer la différence de réflectance entre les bandes les plus sensibles (moyen infrarouge et rouge) et les bandes les moins sensibles (bleu) permettra de maximiser la variation. Tandis que pour les surfaces en eau, la réflectance en bande bleu est plus élevée que la réflectance en bande rouge et moyen infrarouge. Le VSDI varie dans un intervalle de valeur positive, une valeur inférieure à 1 correspond à des zones humides (végétation et sol) et une valeur supérieure à 1 représente des surfaces en eau.

3.4.1.2 Images satellites Spot

Les séries temporelles d'images satellites Spot comptent des images Spot-2, Spot-4 et Spot-5. Afin de caractériser ces images satellites, nous avons considéré les trois bandes spectrales en commun : verte, rouge, proche infrarouge. Nous avons aussi calculé les indice spectraux suivants : l'indice de végétation ajusté au sol (SAVI), l'indice de brillance (BI), l'indice de couleur (CI), indice de différence d'eau normalisé (NDWI II), et Indice de végétation par différence normalisé (NDVI). Ces indices sont introduits pour caractériser la végétation d'une part et les surfaces en eau d'autre part.

— Indice de végétation ajusté au sol (SAVI) (Huete, 1988) : le SAVI est une des variantes de NDVI, il permet de corriger les effets de brillance du sol (sol

clair ou sol foncé) sur la réponse spectrale du couvert végétal, en particulier le couvert végétal peu dense. Cet indice est calculé à partir des bandes Rouge (R) et Proche Infrarouge (PIR) en se basant sur l'équation : $(1 + L)(PIR - R)/(PIR + R + L)$. Le SAVI est calculé de la même façon que le NDVI tout en intégrant un facteur de correction de la luminosité du sol (L).

— Indice de brillance (BI) (Girard et Girard, 2010) : Le BI permet d'estimer le brillance du sol en utilisant la bande Rouge (R) et la bande Proche Infrarouge (PIR), il est par l'équation : $\sqrt{(R^2 + PIR^2)}$. Cet indice permet de discriminer le couvert végétal de sol nu. En effet, la végétation, qu'elle soit verte ou sèche, est souvent plus sombre que le sol sur lequel elle se développe. Il permet aussi de distinguer les différents états pour un même sol nu en fonction de sa teneur en eau. Les plages, par exemple, sont caractérisées par une brillance élevée, tandis que les zones d'eau et les zones humides sont sombres.

— Indice de couleur (CI) : Le CI permet de caractériser le sol, il est calculé à partir des bandes Verte (V) et Rouge (R) par l'équation : $(R - V)/(R + V)$. Plusieurs travaux ont montré la forte corrélation entre les couleurs des sols et leurs réflectances dans les bandes du visible (Escadafal, 1993). La réponse spectrale du sol est influencé par sa composition, ainsi le CI combiné au BI et au NDVI permet de déterminer sa couverture.

— Indice de différence d'eau normalisé (NDWI II) : le NDWI II permet de caractérisé les surfaces en eau, il est dérivé à partir des bandes Verte (V) et Proche Infrarouge (PIR) : $(V - PIR)/(V + PIR)$. La bande verte permet de maximiser la réflectance des surfaces en eau, tandis que la bande proche infrarouge permet de minimiser leur réflectance. La bande proche infrarouge permet aussi de maximiser la réflectance du couvert végétal. Ainsi les valeurs positives de cet indice sont assimilées aux surfaces en eau et les valeurs négatives ou nulles sont assimilées à de la végétation.

— Indice de végétation par différence normalisé (NDVI) : Voir Section 3.4.1.1

3.4.2 Segmentation

L'étape de segmentation vise à partitionner l'image en un ensemble de segments/objets (Blaschke, 2010). Les objets représentent des entités homogène du paysage : parcelle agricole, forêt, etc. Ils sont constitués d'un ensemble de pixels voisins similaires. Les critères de similarité adoptés sont : spectrale, texturale et spatiale. Ils sont définis différemment selon la méthode de segmentation adoptée. Il existe plusieurs méthodes de segmentation dans la littérature. Dans le cadre de notre thèse, nous avons utilisé deux méthodes de segmentation différentes. Les série temporelles d'images satellites Landsat ont été segmentées en appliquent l'algorithme de segmentation multirésolution (Baatz et Schäpe, 2000), tandis que les séries temporelles d'images satellites Spot ont été segmentées en appliquant l'algorithme *mean shift* (Comaniciu et Meer, 2002). Des expérimentations empiriques nous ont

permet d'identifier l'algorithme et les paramètres adéquats pour les images satellites Landsat et Spot utilisées.

3.4.2.1 Image satellites Landsat

Les images Landsat-5 décrivant la Vallée du Libron et la Basse Plaine de l'Aude on été segmentées en utilisant l'algorithme de segmentation multirésolution. Cet algorithme est disponible dans l'outil eCognition Developer 8.8.1. La segmentation est basée sur l'ensemble de tout les descripteurs des images, à savoir les six bandes spectrales et les trois indices calculés. Chaque image a été segmentée indépendamment des autres images de la série temporelle. Le Tableau 3.3 reporte les résultats de l'étape de segmentation. Nous notons que 3373 objets ont été identifiés sur la Basse Plaine de l'Aude, tandis que sur la Vallée du Libron, seulement 1218 objets ont été extrait. En effet, la superficie du premier site est plus grande que la superficie du deuxième site.

	Pixel	Objet
Basse Plaine de l'Aude	53 782	3 373
Vallée du Libron	18 394	1 218

TABLE 3.3 – Le nombre d'objets résultant de la segmentation des STIS annuelles décrivant la Basse Plaine de l'Aude et la Vallée du Libron.

3.4.2.2 Image satellites Spot

Les séries temporelle pluriannuelles décrivant la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas et le Pic Saint Loup ont été segmentées en utilisant l'algorithme mean shift disponible sous le logiciel open source de traitement d'image satellite QGIS. L'information spectral utilisée lors de la segmentation compte les trois bandes spectrales brutes et l'indice de végétation par différence normalisé. De même chaque image a été segmentée indépendamment des autres images de la série. Les résultats obtenus sont reportés dans le Tableau 3.4. La MMCA représente le site le plus grand parmi nos sites d'étude ce qui explique le grand nombre d'objet détectés lors de la segmentation. Le nombre d'objets détectés dans la Basse Plaine de l'Aude quant à lui reflète l'hétérogénéité de l'occupation du sol sur ce site. Tandis que Pic Saint Loup est caractérisé par un paysage homogène, ce qui explique le nombre d'objets résultant de sa segmentation.

3.5 Données de références

Chacune des séries temporelles a été analysée par un expert du terrain afin d'identifier la couverture de sol des différents sites d'étude. Dans ce qui suit, nous présentons les cartes d'occupation du sol de chaque site d'étude.

	Pixel	Objet
Basse Plaine de l'Aude	116358	13075
Montagne de la Moure et Causse d'Aumelas	237886	17288
Pic Saint Loup	124312	12184

TABLE 3.4 – Le nombre d'objets résultant de la segmentation des STIS pluriannuelles décrivant la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas et le Pic Saint Loup.

3.5.1 Images satellites Landsat

La Basse Plaine de l'Aude et la Vallée du Libron sont décrites au travers de séries temporelles d'images satellites issues du capteur Landsat-5. Leur analyse par un expert du terrain a permis d'identifier onze types de couverture du sol sur le premier site et neuf types de couverture du sol sur le deuxième. L'expert a aussi identifié quelques zones inclassables sur les deux sites. La Figure 3.12 décrit la nomenclature déterminée pour chacun des deux sites. La Basse Plaine de l'Aude est caractérisée par la diversité de son occupation du sol, on y est trouve : des couverts végétaux, des *zones humides*, des *étangs* et aussi des zones littorales (*plage*). La Vallée du Libron est prédominée par le couvert végétal (*forêt*, *culture*, etc), on y est trouve aussi quelques zones urbanisées et *espaces artificialisés*.

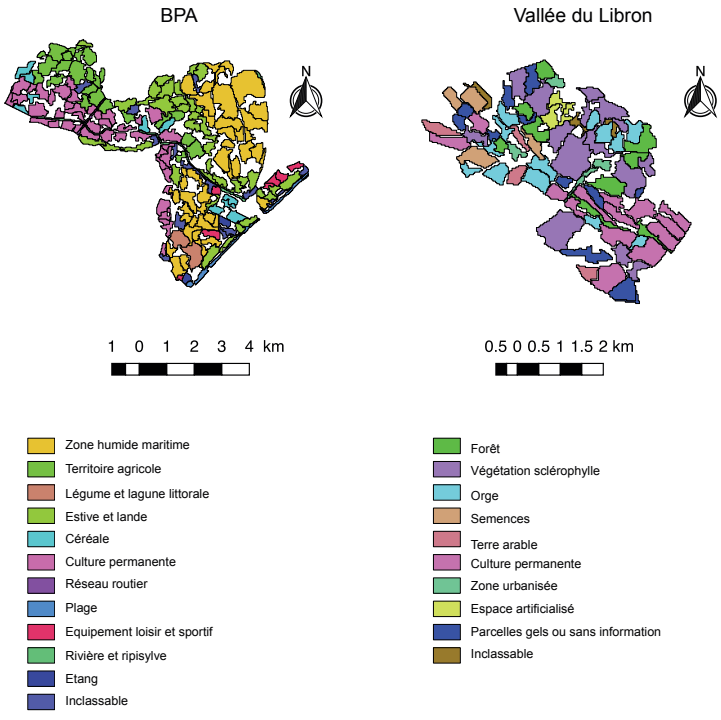


FIGURE 3.12 – La répartition des classes de couverture du sol identifiées dans les deux sites : la Basse Plaine de l'Aude et la Vallée du Libron.

3.5.2 Image satellites Spot

Les série temporelles d'images satellites décrivent la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas et le Pic Saint Loup sont issue du capteur Spot. La Figure 3.14 décrit les différent types d'occupation du sol de chaque site. La Basse Plaine de l'Aude compte six classes dont les couverts végétaux, les *zones humides*, les *surface en eau* et les *plages*. La MMCA est caractérisée par un paysage naturel prédominé par la végétation répartie en cinq classes : *forêt*, *culture*, *espace ouvert avec peu ou sans végétation*, *milieu à végétation arbustive* et *vignoble*. Le Pic Saint Loup est lui aussi prédominé par la végétation, il compte six classes dont *culture*, *forêt*, *espace couvert avec peu ou sans végétation*, *milieu a végétation arbustive*, *vignoble* et quelques *espaces artificialisés*.

Les trois sites partagent certaines occupations du sol comme illustré dans la Figure 3.13. Les classes en commun sont au nombre de trois dont : *culture*, *milieu à végétation arbustive*, et *vignoble*. Le Pic Saint Loup et la Montagne de la Moure et Causse d'Aumelas partagent aussi deux occupation du sol qui sont : *espace ouvert avec peu ou sans végétation* et *forêt*.

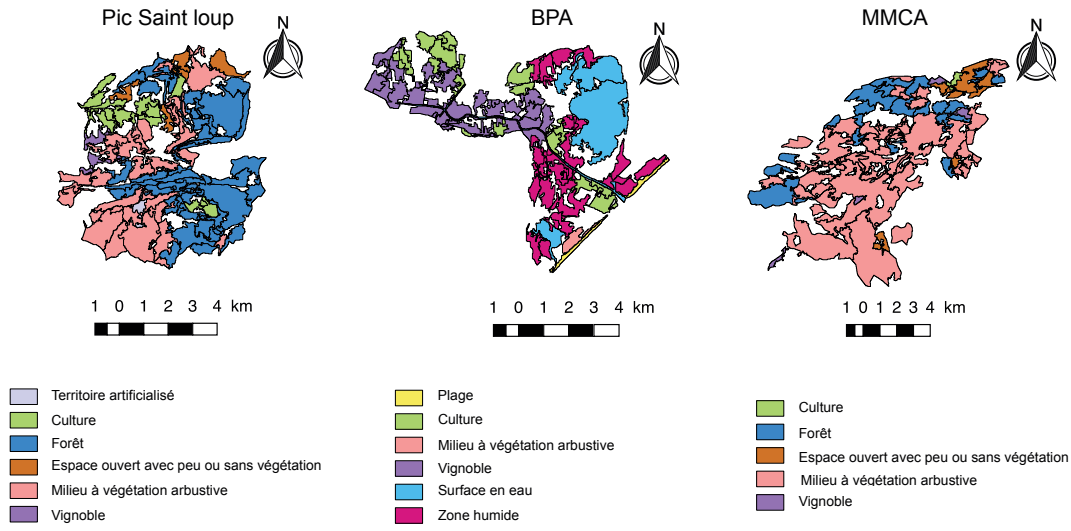


FIGURE 3.13 – La répartition des classes de couverture du sol identifiées dans les trois sites. Les classes en commun sont : *culture*, *milieu à végétation arbustive*, et *vignoble*.

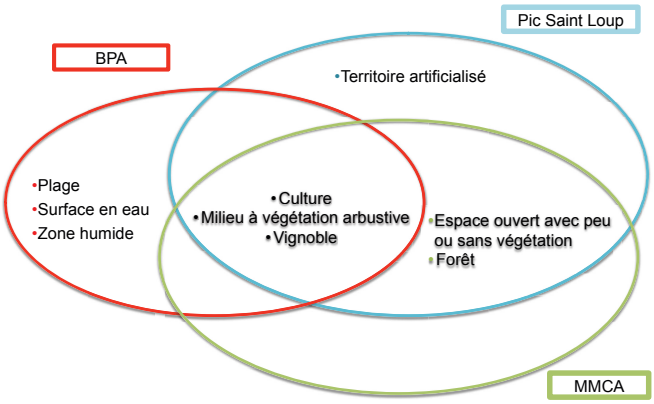


FIGURE 3.14 – Les classes de couverture du sol identifiées dans les trois sites : la Basse Plaine de l'Aude, la Montagne de la Moure et Causse d'Aumelas et le Pic Saint Loup . Les classes en commun sont : *culture*, *milieu à végétation arbustive*, et *vignoble*.

3.6 Conclusion

Dans cette thèse, nous nous intéressons au suivi de l'évolution des zones naturelles et semi-naturelles par imagerie satellite. Afin d'atteindre cette objectif, nous allons adopter une analyse d'images satellites orientée object en s'appuyant sur les méthodes d'apprentissage non supervisées et semi-supervisées. Afin de valider nos approches, nous avons sélectionné quatre sites d'étude : la Basse Plaine de l'Aude, la Vallée du Libron, MMCA et le Pic Saint Loup, localisés au sud de la France.

Dans ce chapitre, nous avons d'abord présenté ces quatre sites étude ainsi que les images satellites décrivant chacun d'eux. Ces images satellite sont issues de capteurs différents (Spot et Landsat) possédant des caractéristiques spécifiques. Nous avons ensuite décrit les étapes de pré-traitement de ces images. Enfin, nous avons analysé la couverture du sol des quatre sites.

Dans les chapitres suivant (Chapitres 4, 5 et 6), nous allons présenter les différentes méthodes d'analyse de séries temporelles d'image satellite proposées ainsi que leur application aux données décrites dans ce chapitre.

Chapitre 4

Analyse non supervisée de séries temporelles d'images satellites

Sommaire

4.1 Introduction

L'objective de notre thèse est l'étude des dynamiques des habitats naturels et semi-naturels par télédétection. Dans le chapitre 2, nous avons présenté les principales approches d'analyse d'images satellites. La littérature foisonne de méthodes de cartographie du sol. On distingues les approches orientées pixel des approche orientées objet.

Les premières approches d'analyse se basent sur les pixels qui sont la plus petite unité de l'image, cette unité n'apporte pas une analyse sémantique du paysage. Tandis que les approches orientées objet traitent des images segmentées sur les quelles des unités structurales du paysage ont été identifiées.

Nous travaux s'inscrivent dans ce contexte. Dans ce chapitre, nous présentons la première de nos deux propositions qui ont pour objective l'analyse non supervisée des dynamiques des habitats naturels et semi-naturels. L'objectif général de notre proposition est l'exploitation des séries temporelles d'images satellites segmentées pour une analyse orientée objet (OBIA- Object Oriented Image Analysis) des évolutions des entités spatio-temporelles pour l'identification et la mise en évidence des patrons d'évolution.

Les travaux d'analyse de STIS, déjà présents dans le littérature traitent de la détection de changements en comparant deux images satellites segmentées. La problématique principale qui en ressort est l'alignement des objets dans une série temporelle. En effet, quand il s'agit d'aligner des pixels, il suffit de superposer les images satellites. Il est plus fastidieux pour l'alignement d'objets car il n'y pas de correspondance un à un entre les images. Afin de remédier a cette problématique, nous proposons la méthode nommée *Evolution Graph Based Clustering*.

Cette méthode compte trois étapes : (i) la détection des entités spatio-temporelles (objets de référence), (ii) la construction des graphes d'évolutions et (iii) le clustering des graphes d'évolutions.

Les entités spatio-temporelles représentent les objets à analyser, les graphes d'évolutions permettent quant à eux d'illustrer l'évolution des entités spatio-temporelles et finalement le clustering des graphes permet d'organiser et de mettre en évidence des objets qui évoluent de la même façon mais aussi des patrons d'évolution.

La représentation des évolutions des objets par graphe a été introduite par (Guttler et collab., 2017). Nous avons adopté la même représentation. Nous avons proposé une méthode d'analyse adaptée aux graphes. Afin de valider notre travail et de le situer dans l'état de l'art, nous l'avons comparé aux méthodes d'analyse basées pixel.

Dans ce chapitre, nous décrivons les différents étapes de la méthode proposée pour l'analyse des séries temporelles d'images satellites orientée objet (section 4.2). Nous présentons aussi les résultats expérimentaux obtenus (section 4.3).

4.2 Analyse non supervisée des STIS

Dans cette section, nous présentons le processus général de notre méthodologie d'analyse des séries temporelles d'images satellites puis, nous décrivons en détail chacune de ses étapes.

4.2.1 Vue globale de l'approche

La Figure 4.1 décrit le processus global de notre méthode d'analyse des séries temporelles d'images satellites. Cette méthode compte trois étapes : la détection des objets de référence, la construction des graphes et le clustering des graphes. Étant donnée une zone d'étude à analyser, nous considérons une série temporelles décrivant cette zone ainsi que les données de segmentation qui lui correspondent. Tout d'abord, les objets résultants de la segmentation de la série temporelle sont filtrés afin de sélectionner un sous-ensemble d'objets nommé, **objets de référence** (Figure 4.1 (Étape 1)). Ensuite, pour chaque objet de référence, est construit un graphe nommé, **graphe d'évolution**. Le graphe d'évolution permet de décrire l'évolution de l'objet de référence, au travers du temps, en terme de couverture du sol. Il est construit en considérant les objets couvrant la même étendue spatiale que l'objet de référence dans toute les images satellites de la série (Figure 4.1 (Étape 2)). Pour chaque objet de référence sélectionné lors de la première étape, il est construit un graphe d'évolution. Une fois tout les graphes d'évolutions construits, ils sont analysés et regroupés dans des groupes homogènes. Cette étape est nommée clustering des graphes (Figure 4.1 (Étape 3)). Elle permet, via des algorithmes de clustering de mettre en évidence les graphes représentant des objets de référence qui évoluent similairement. Cette dernière étape a pour objectif d'organiser et d'identifier des patrons d'évolutions dans nos zones d'étude.

4.2.2 Détection des entités spatio-temporelles

Les entités spatio-temporelles correspondent aux objets de référence d'intérêt à analyser. Un objet de référence peut représenter, par exemple, une parcelle agricole. Dans ce cas là, le phénomène à observer serait le cycle de croissance des plantations. Les objets de référence constituent un sous-ensemble de $\mathcal{O}$ c'est-à-dire les objets issue de la segmentation de la série temporelle sont analysés afin d'en sélectionner un sous ensemble dit, objets de référence ($RefObjs$). Ce processus de détection se base sur l'hypothèse suivante : un phénomène à observer est défini par une étendue spatiale (nombre de pixel qu'il couvre) dans chaque image de la série. La nature du phénomène influence sur son étendue spatiale qui peut changer de taille. Nous présentons comme exemple l'évolution d'un lac. En saison d'hiver, le lac est rempli d'eau tandis qu'en saison d'été, le lac s'assèche et l'eau recouvre alors moins de surface. Afin de pouvoir observer ce phénomène, l'objet de référence qui doit être considéré est l'étendue du lac en période d'hiver. Cela permet d'avoir une information

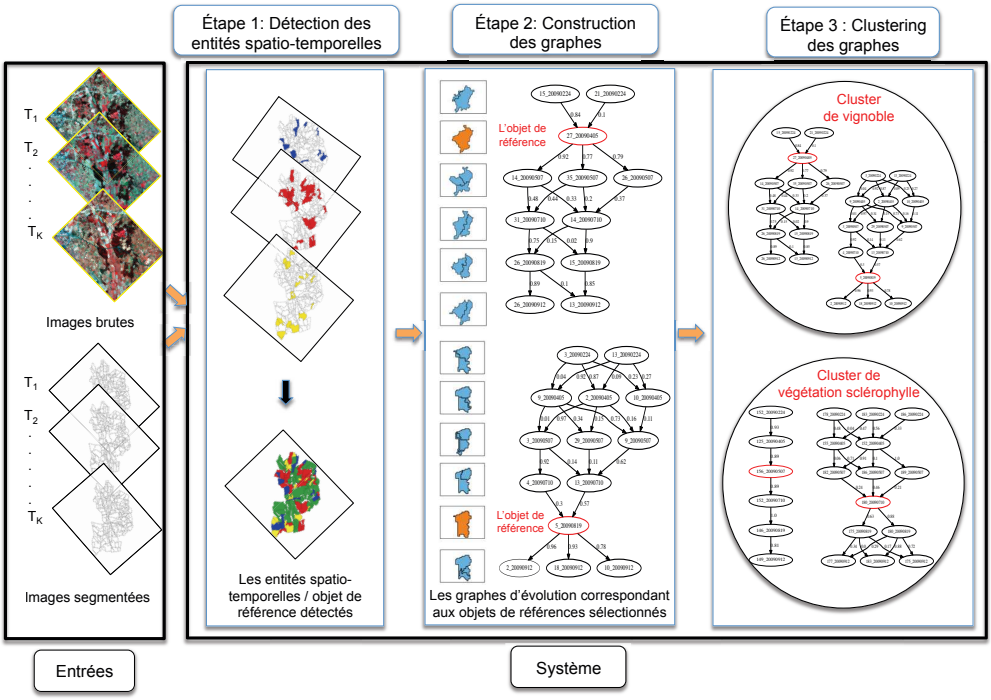


FIGURE 4.1 – Schéma général décrivant les différentes étapes du processus d'extraction et d'analyse des entités spatio-temporelles (objets de référence).

complète pour observer le processus d'assèchement du lac. Ainsi nous sélectionnons les objets les plus grands de la série temporelle.

La détection des objets de référence est traité comme un problème de recouvrement. Les objets de référence sont sélectionnés de manière à couvrir le maximum possible de la zone d'étude en minimisant leur chevauchement. En effet, nous considérons des objets appartenant à toute la série, deux objets appartenant à deux images différentes peuvent ainsi se chevaucher.

L'algorithme 3 décrit la procédure de détection des objets de référence qui reçoit en entrée deux paramètres : $CandidateObj$ et α. Le premier paramètre représente l'ensemble des objets candidats, celui-ci est un sous-ensemble de $\mathcal{O}$, pour chaque pixel de la zone d'étude on sélectionne l'objet le plus grand de la série le couvrant. Tandis que le deuxième définit un seuil de chevauchement entre les objets de référence ($RefObj$). Ce processus est itératif, à chaque itération tous les objets sont pondérés (ligne 8-12) et l'objet dont le poids est le plus élevé est sélectionné (ligne 15-17) et ajouter à l'ensemble des objets de référence (ligne20-21), tandis que les objets dont le poids est inférieur à α sont écartés (ligne 13-14).

Les poids des objets définissent leur pertinence, ou non, à être analysé. Le pro-

Algorithme 3 : DétectionRefObj($CandidateObj$**,** α**)**

1: Entrées : $CandidateObj$
2: α
3: Sorties : $RefObj$
4: $RefObj = \emptyset$
5: $poids = 0$
6: **tant que** $|Objets| > 0$ **faire**
7: $obj = NULL$
8: $max\_poids = -1$
9: **pour tout** $o \in Objets$ **faire**
10: //Pondération des objets
11: **si** $Pix(o) \bigcap Pix(RefObj) = \emptyset$ **alors**
12: $poids = |Pix(o)|$
13: **sinon**
14: $poids = \frac{|Pix(o) - Pix(RefObj)|}{|Pix(o)|}$
15: **fin si**
16: //Elimination des objets dont le poids est inférieur au seuil fixé
17: **si** $poids < \alpha$ **alors**
18: $Objets = Object$ - o
19: //Sélection de l'objet dont le poids est maximal
20: **sinon si** $Poids > max\_poids$ **alors**
21: $max\_poids = poids$
22: $obj = o$
23: **fin si**
24: **fin pour**
25: //Ajout de l'objet sélectionné à l'ensemble des objets de référence
26: **si** $obj \neq NULL$ **alors**
27: $RefObj = RefObj \cup obj$
28: **fin si**
29: **fin tant que**
30: **retourner** $RefObj$

cessus de pondération se base sur l'étendue spatiale de l'objet (nombre de pixel qu'il couvre) ainsi que sur la zone couverte (les pixel couvert) par les objets de référence déjà sélectionnés. Ce poids est égal à la taille du l'objet si celui-ci ne chevauche pas l'ensemble des objets de référence déjà sélectionnés (ligne 8-9). Si au contraire l'objet chevauche les objets de référence déjà sélectionnés, son poids est égal au ratio entre le nombre de pixel de l'objet non couvert par l'ensemble des objets de référence déjà sélectionnés, et sa taille (ligne 11).

En résumé, le processus de détection des objets de référence commence d'abord par sélectionner tous les objets les plus grands de la série qui ne se chevauchent pas, puis sélectionne ceux dont le chevauchement est inférieur à α. Ce processus se termine lorsque les objets de référence couvrent toute la zone d'étude ou bien le taux chevauchement entre les objets ne satisfait plus le seuil α.

4.2.3 Construction des graphes d'évolution

Une fois tous les objets de référence sélectionnés, l'étape suivant est l'observation de leur dynamique comme par exemple, l'évolution d'une parcelle agricole qui compte la saison de levé de culture et la saison de récolte.

Pour chaque objet de référence o^* sélectionné précédemment, on construit un graphe d'évolution noté G_{o^*}. Un graphe d'évolution est un graphe acyclique orienté,

il est défini par un ensemble de nœuds et un ensemble d'arcs, $G_{o^*} = (V_{o^*}, E_{o^*})$. L'ensemble de nœuds V_{o^*} correspond aux objets de la série temporelle quand à l'ensemble d'arcs E_{o^*}, il correspond au lien entre les objets.

La construction des graphes d'évolution se fait en trois étapes : (i) La définition de l'ensemble de nœuds, (ii) l'organisation des nœuds en couches et (iii) la définition de l'ensemble des arcs.

La premier étape consiste à sélectionner les objets qui correspondent aux nœuds du graphe. Pour cela, nous considérons l'étendue spatiale de l'objet de référence. Nous sélectionnons sur chaque image de la série les objets qui recouvrent les même pixels que l'objet de référence. Étant donné que chaque image est segmentée indépendamment, il y a pas de correspondance un à un entre les objets des images. Ainsi pour chaque objet de référence, nous identifions l'ensemble d'objets le plus optimal qui couvre les même pixels, c'est-à-dire le plus petit ensemble d'objets qui couvre la même étendue spatiale en évitant de perdre de l'information (les objets couvrent moins de pixel que l'objet de référence) et d'inclure du bruit (les objets couvrent plus de pixel que l'objet de référence). Les objets sélectionnés doivent répondre à deux critères : un nombre suffisant de pixels de l'objet se chevauchent avec l'objet référence ou l'objet couvre un nombre suffisant de pixels de l'objet de référence. Ces deux critères sont formalisés au travers de deux seuils, σ_1 et σ_2 respectivement. L'équation 4.1 définit l'ensemble de nœuds du graphe G_{o^*}.

$$V_{o^*} = \{o | o \in O, \frac{|Pix(o^*) \cap Pix(o)|}{|Pix(o)|} \geq \sigma_1 \; ou \; \frac{|Pix(o^*) \cap Pix(o)|}{|Pix(o^*)|} \geq \sigma_2\} \tag{4.1}$$

Où :

— O représente l'ensemble des objets ;

— o est un objet de O ;

— o^* est un objet de référence ;

— $Pix(o)$ représente l'ensemble de pixels couvert par l'objet o.

Ainsi pour chaque image de la série, on sélectionne un groupe d'objets. Les groupes sont ensuite ordonnés dans le temps en se basant sur les dates d'acquisition des images satellites qui leur correspondent. Les objets appartenant à deux images successives sont ensuite reliés, un arc est défini entre chaque pair d'objets qui se chevauchent. Le nœud source de l'arc est définit par l'objet dont la date d'acquisition est la plus antérieure tandis que le nœud destination est définit par l'objet dont la date d'acquisition est la plus postérieure. L'équation 4.2 définit l'ensemble des arcs du graphe G_{o^*}.

$$E_{o^*} = \{(o^i, o^j) | o^i \in O_{T_l} \cap V_{o^*}, o^j \in O_{T_{l+1}} \cap V_{o^*}, Pix(o^i) \cap Pix(o^j) \neq \emptyset\} \tag{4.2}$$

Où :

— O_{T_l} and $O_{T_{l+1}}$ représentent l'ensemble des objets correspondant aux images I_{T_l} et $I_{T_{t+1}}$ respectivement ;

— o^i and o^j sont des objets des images I_{T_l} et $I_{T_{l+1}}$ respectivement ;

— $Pix(o^i)$ représente l'ensemble des pixels couverts par l'objet o^i ;

La figure 4.2 illustre un exemple de graphe d'évolution correspondant à un objet de référence de la Vallée du Libron. Ce graphe d'évolution se compose de six groupes d'objets correspondant aux six images de la série temporelle d'images satellites annuelle décrivant la Vallée du Libron. Les groupes d'objets sont ordonnés en couche de gauche à droite selon la ligne du temps. L'objet de référence correspond au nœud orange. Les objets appartenant à la même image ne sont pas reliés entre eux, seuls les objets de deux images différentes et dont les dates d'acquisition sont successives sont reliés.

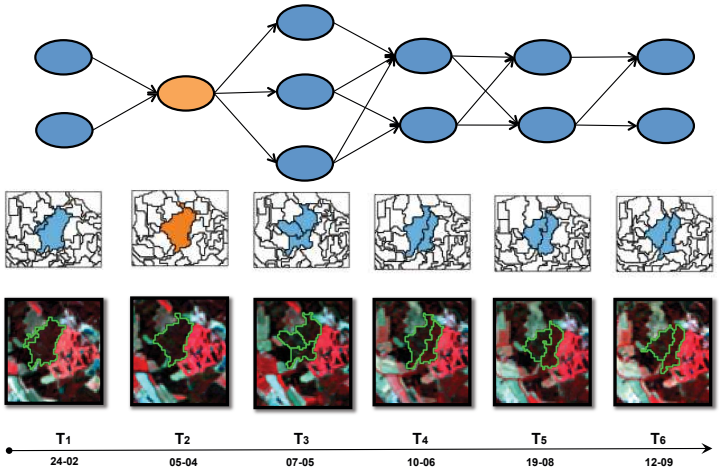


FIGURE 4.2 – Exemple d'un graphe d'évolution représentant l'évolution d'une zone couverte par de la *végétation sclérophylle* identifiée dans la deuxième image T_2 (l'objet orange).

4.2.4 Clustering des graphes d'évolution

Le clustering des graphes d'évolutions représente la dernière étape de notre méthodologie d'analyse des dynamiques des entités spatio-temporelles (objets de référence). Les graphes d'évolution sont transformés en synopsis. Les synopsis sont définis comme une séquence d'objets, ils sont utilisés pour estimer les distances entre les graphes d'évolutions qui leur correspondent.

Les synopsis permettent de résumer l'information spectrale des graphes d'évolutions, ils correspondent une séquence d'objets. Les objets qui composent les synopsis résultent de l'agrégation pondérée des objets des graphes d'évolutions. la figure 4.3 illustre la méthode de construction des synopsis.

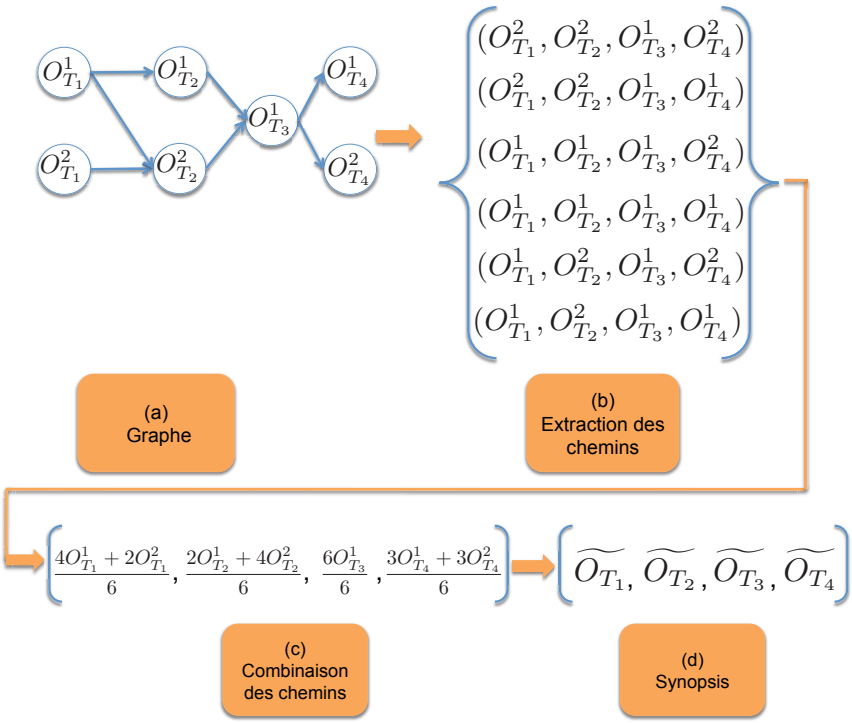


FIGURE 4.3 – Procédure de génération des synopsis à partir des graphes d'évolution.

La premier étape dans le processus de transformation du graphe en synopsis est l'extraction de ses chemins. Un chemin est défini comme une séquence ordonnée d'objets adjacents (lié par des arcs). Tel que nous définissons les chemins, le premier objet de la séquence, noté s, doit appartenir à la première image de la série temporelle c'est-à-dire, $s = o^i$ tel que $o^i \in O_{T_1} \cap V_{o^*}$. Quant au dernier objet, noté v, de ce chemin, il doit appartenir à la dernier image de la série temporelle, c'est-à-dire, $v = o^j$ tel que $o^j \in O_{T_K} \cap V_{o^*}$. L'ensemble des chemins extraits du graphe d'évolution G_{o^*} est noté, $\mathcal{P}_{G_{o^*}}$.

En considérant la structure particulière des graphes d'évolution, les chemins qui les composent contiennent chacun T objets, c'est-à-dire un objet par image. Quant à l'objet de référence, il est contenu dans tous les chemins.

Une fois l'ensemble des chemins $\mathcal{P}_{G_{o^*}}$ extraits pour chaque graphe d'évolution, les synopsis sont générés. Les synopsis sont définis comme une séquence de K objets, un objet $\widetilde{O_{T_l}}$ par image I_{T_l}. Chaque objet $\widetilde{O_{T_l}}$ est le résultat d'une aggrégation pondérée des valeurs radiométriques des objets appartenant à l'image I_{T_l} de chacun des chemins.

L'agrégation pondérée des objets permet de donner plus d'importance aux objets les plus influents dans les graphes d'évolution. Un objet est considéré comme étant influent s'il concerne plusieurs chemins du graphe. L'étape d'agrégation (Fig. 4.3(c)) calcule la moyenne des valeurs radiométriques des objets en associant à chaque ob-

jet un poids égal au nombre de chemins dans lesquels il apparait. Chaque objet est sommé autant de fois qu'il apparait dans un des chemins du graphe d'évolution. Considérant l'exemple donné par la figure 4.3, nous considérons un graphe d'évolution qui se compose de quatre groupe d'objets (issues d'une série temporelle qui se compose de quatre images). Nous pouvons observer que l'objet $o^1_{T_1}$ apparait dans quatre chemins tandis que l'objet $o^2_{T_1}$ apparait dans deux chemins. Ainsi l'objet $o^1_{T_1}$ est plus important que l'objet $o^2_{T_1}$. Les deux objet sont donc pondérés différemment, l'objet $o^1_{T_1}$ a un poids égal à quatre et l'objet $o^2_{T_1}$ a un poids égal à deux. L'équation 4.3 définit le calcul des objets des synopsis.

$$Info(\widetilde{O_{T_l}}) = \frac{\sum_{o_{T_l} \in \mathcal{P}_{G_{o^*}}} Info(o_{T_l})}{|\mathcal{P}_{G_{o^*}}|} \tag{4.3}$$

Chaque graphe d'évolution est alors assimilé à un synopsis dont l'information radiométrique lui est équivalente. Ainsi la distance entre chaque paire de synopsis est équivalente à la distance entre leur graphes d'évolutions respectives. Considérant deux graphes d'évolution G_1 and G_2, nous générons d'abord leur synopsis respectif syn_1 et syn_2. Pour estimer la distance entre les deux synopsis syn_1 et syn_2, on calcule la distance entre leurs objets comme défini par l'équation 4.4.

$$dist_s(syn_1, syn_2) = \frac{\sum_{l=1}^{|syn_1|} dist(syn_1[T_l], syn_2[T_l])}{|syn_1|} \tag{4.4}$$

La distance entre deux objets est donnée par la distance euclidienne entre leur vecteur d'information radiométrique comme définit par l'équation 4.5.

$$dist(\widetilde{O_{T_l}}, \widetilde{O_{T'_l}}) = ||Info(\widetilde{O_{T_l}}) - Info(\widetilde{O_{T'_l}})||_2 \tag{4.5}$$

Pour chaque paire de synopsis, nous calculons une distance équivalente à la distance entre leur graphe d'évolution respectif. Cette procédure nous permet ainsi de construire une matrice de distance contenant la distance entre chaque paire de graphes d'évolution. Cette matrice est ensuite utilisée pour détecter des groupes de graphes d'évolution homogènes qui correspondent à des entités spatio-temporelles dont les dynamiques sont similaires. Le partitionnement des graphes d'évolution, dit clustering des graphes d'évolution, est réalisé en appliquant des algorithmes d'apprentissage non supervisé (Tan et collab., 2005)(algorithme de clustering : clustering Hiérarchique, K-Means, clustering Spectral , DBSCAN, etc.) sur la matrice de distance indépendamment des données des graphes d'évolution.

4.3 Expérimentations

Afin de valider notre méthode d'analyse des évolutions des entités spatio-temporelles, nous avons mené des expérimentations sur les deux séries temporelles d'images satellites annuelles décrivant la Vallée du Libron et la Basse Plaine de

l'Aude (Section 3.3.1). Nous présenterons, dans ce qui suit, les résultats obtenus. Nous décrivons tout d'abord l'approche adoptée pour le choix des paramètres. Ensuite nous présentons les résultats quantitatifs issus de la comparaison de notre méthode à celles de l'état de l'art. Enfin, nous analysons quelques exemples de graphes d'évolution générés sur les deux sites.

4.3.1 Sélection des paramètres

La méthode d'analyse des évolutions des entités spatio-temporelles intègre trois paramètres : α, σ_1, et σ_2. Ces trois paramètres varient dans l'intervalle $[0, 1]$.

Le paramètre *alpha* intervient dans l'étape de détection des objets de référence et permet de sélectionner l'ensemble des objets de référence le plus pertinent c'est-à-dire, l'ensemble des objets de référence qui permet de couvrir le maximum de la zone d'étude en minimisant le chevauchement des objets.

Les deux paramètres, σ_1 et σ_2, interviennent dans l'étape de construction des graphes d'évolutions. Ils permettent de sélectionner les nœuds du graphe.

Les valeur de ces trois paramètres sont fixées automatiquement en se basant sur deux indicateurs : la couverture des graphes d'évolution et le chevauchement des graphes d'évolution :

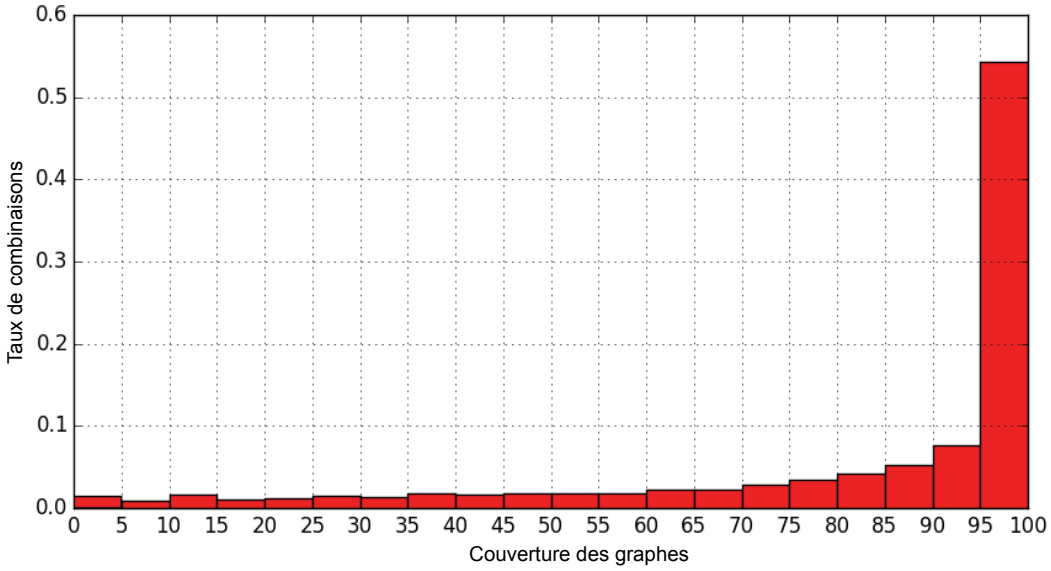
— La couvertures des graphes d'évolution : la couverture d'un graphe d'évolution est défini par la surface couverte par tous les objets du graphe d'évolution toutes images confondues. La couvertures des graphes d'évolution est l'union des couvertures de chaque graphe d'évolution.

— Le chevauchement des graphes d'évolution : le chevauchement des graphes d'évolution correspond au chevauchement entre la couverture des graphes d'évolution. Ce chevauchement est défini par la surface couverte par au moins deux graphes d'évolution ou plus.

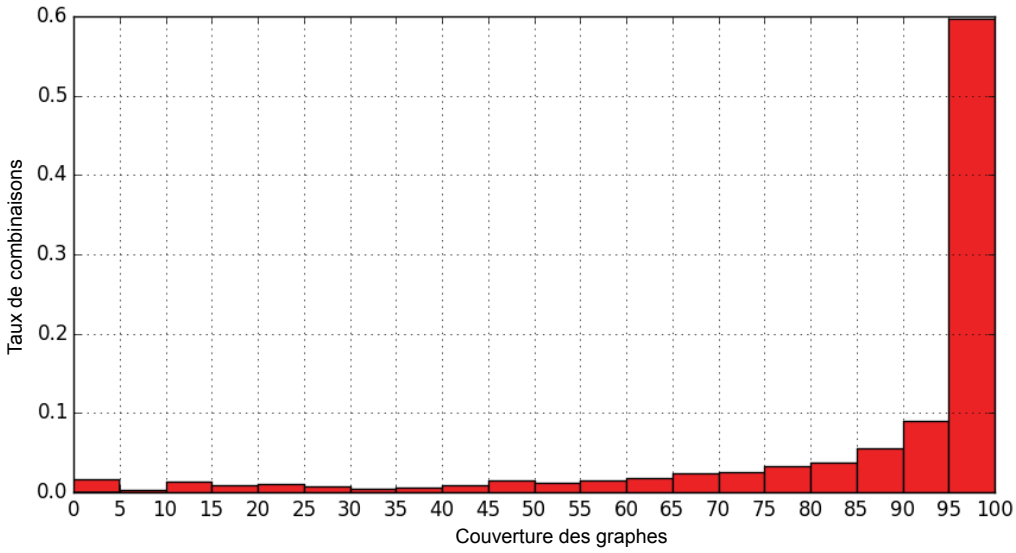
Nous avons proposé une approche heuristique afin de sélectionner les valeurs de ces trois paramètres. Nous avons d'abord fait varier les trois paramètres dans l'intervalle $[0, 1]$ avec un pas de 0.5. Ensuite, pour chaque combinaison de valeurs possibles pour les trois paramètres, nous avons généré l'ensemble des graphes d'évolution lui correspondant. Nous avons enfin rapporté le taux de couverture des graphes d'évolution ainsi que leur taux de chevauchement.

La figure 4.4 présente le taux de couverture des graphes d'évolutions construit pour les différentes combinaisons possibles des paramètres sur les deux zones d'étude : la Basse Plaine de l'Aude (figure 4.4 (a)) et la Vallée du Libron (figure 4.4 (a)).

Nous observons sur la figure 4.4 que plus de 50% des combinaisons génèrent un ensemble de graphes d'évolution qui couvre plus de 95% de la zone d'étude, aussi bien pour la Basse Plaine de l'Aude que pour la Vallée du Libron. En s'appuyant sur ces résultats, nous avons fixé un seuil τ pour la couverture des graphes égal à 95%.



(a)



(b)

FIGURE 4.4 – Taux de couverture des graphes d'évolution générés par les différentes combinaisons sur les trois sites d'étude (a) la Basse Plaine de l'Aude et la (b) Vallée du Libron .

Nous avons sélectionné toutes les combinaisons qui permettent de générer un ensemble de graphes d'évolution couvrant 95% au plus de la zone d'étude. Parmi ces combinaisons, nous avons ensuite sélectionné la combinaison qui correspond à l'ensemble des graphes d'évolutions dont le chevauchement est minimal. Les valeurs

définies sur les deux zones d'étude sont reportées dans le tableau 4.1.

	α	σ_1	σ_2	taux de couverture des graphes	taux de chevauchement des graphes
Basse Plaine de l'Aude	0.8	0.4	0.95	95.5	34.7
Vallée du Libron	0.7	0.45	0.85	95.2	24

TABLE 4.1 – Valeurs des paramètres α, σ_1 et σ_2 définis sur la basse plaine d'aude et la Vallée du Libron.

Les valeurs définies pour ces paramètres ont permis de construire 81 graphes d'évolution sur la Vallées du Libron couvant 95.2% de ce site et 194 sur la Basse Plaine de l'Aude avec un taux de couverture de 95.5%.

4.3.2 Protocole d'évaluation quantitative

Afin d'évaluer notre méthode, nous avons interrogé un expert du terrain qui établi une classification de notre zone d'étude. Une fois les graphes d'évolution construits sur chacun des sites d'étude, nous avons fourni les objets de de référence à l'expert qui les a classés dans différentes classes. L'expert du terrain a défini onze classes pour les objets de référence de la Basse Plaine de l'Aude et neuf classes pour ceux de la Vallée du Libron.

Nous avons utilisé cette classification experte pour valider nous résultats automatique. Le critère d'évaluation adopté est la similarité entre le partitionnement automatique et le partitionnement de l'expert, plus les deux partitionnements sont proches meilleure sont nos résultats. Afin d'évaluer la similarité entre les deux partitionnements, nous nous somme appuyés sur les Indices de qualité externes (Section 2.2.4.2).

En outre, nous avons aussi évalué les performances de la notre méthode, nommée *EGraphClustering* (Evolution Graph Based Clustering), à ceux proposées dans l'état de l'art. Nous avons comparé notre méthode qu'est basée objet aux méthode traditionnelle basée pixel. Nous avons considéré trois méthodes :

— *Pixel-Based Clustering* : La première méthode exploite des images satellites brutes pour constituant des séries temporelles de pixels à analyser. Une série temporelle de pixels est une séquence ordonnée de valeurs radiométriques associées au pixel observé dans chaque image de la série temporelle.

— *RObject-Based Clusteing* : La deuxième approche se base sur une analyse par objet et s'appuie uniquement sur l'information radiométrique des objets de référence en ignorant la dimension temporelle des données

— *Pixel-Object-Based Clustering* : Cette dernière méthode PETITJEAN et collab. (2012b) combine l'analyse objet a l'analyse pixel. elle définit des série temporelle de pixels comptant l'information radiométrique du pixel ainsi que celle de son objet dans chaque image de la série temporelle.

L'objectif de la comparaison est double, Nous considéreront d'abord la pertinence de l'information temporelle pour le clustering, en comparant les résultats du clustering des objets de référence aux résultats du clustering des graphes d'évolutions (*EGraphClustering* vs. *RObject-Based Clustering*). Puis nous évaluons la pertinence de l'analyse par objet par rapport à l'analyse par pixel, nous combinons la dimension temporelle et spatiale des données en utilisant la représentation par graphe des entités spatio-temporelles (*EGraphClustering* vs. *Pixel-Based Clustering* et *Pixel-Object-Based Clustering*).

Nous proposons dans ce chapitre une approche d'analyse orientée objet. Ainsi, il serai plus judicieux de se comparer à d'autre approches qui sont elles aussi orientées objet. Cependant, à notre connaissance il n'existe pas dans la littérature de méthode d'analyse de séries temporelles orientée objet.

Afin d'identifier des groupes d'évolutions homogènes, nous avons analysé les graphes d'évolutions à l'aide des algorithmes de clustering. Nous avons utilisé pour cela deux algorithmes différents : l'algorithme de clustering spectral et l'algorithme de clustering hiérarchique (Section 2.2.2). En ce qui concerne la validation des résultats, deux indices de validation, les plus utilisés dans la littérature, ont été employés à savoir l'Indice de Rand Ajusté (ARI) et l'Information Mutuel Normalisée (NMI) (Section 2.2.4.2).

4.3.3 Résultats de l'évaluation quantitative

Les tableaux 4.2 et 4.3 rapportent les résultats expérimentaux sur la Basse Plaine de l'Aude et la Vallées du Libron respectivement c'est à dire, les valeurs du ARI et NMI pour chacune des quatre approches combinées au clustering spectral ainsi qu'au clustering hiérarchique. Ces résultats montrent que l'approche par objet de référence (*RObject-Based Clustering*) est la moins performante des quatre approches soulignant ainsi l'importance de l'information temporelle. Considérer la manière dont les entités spatio-temporelles évoluent au cours du temps permet une meilleure discrimination entre leur classes.

En ce qui concerne les approches basées pixel (*Pixel-Based Clustering* et *Pixel-Object-Based Clustering*) par comparaison à notre approche (*EGraphClustering*), elles sont aussi moins performantes. La méthode *EGraphClustering* sur les deux zones d'études donne des résultats compétitifs par rapport aux autres méthodes. Sur la Vallée du Libron, on obtient les meilleurs scores NMI et ARI en utilisant les deux algorithmes spectral et hiérarchique. Tandis que pour la Basse Plaine de l'Aude, nous obtenons le meilleur score ARI, et des score NMI proches pour toutes les méthodes.

Les séries temporelles d'images satellites diffèrent des séries temporelles standards. Généralement les séries temporelles standards sont fortement caractérisées par la dimension temporelle des données tandis que les STIS sont autant caractérisées par la dimension temporelle que spatiale. Considérer les deux dimensions nous permet de discriminer les différentes évolutions des entités spatio-temporelles. Les

graphes d'évolution nous ont permis de modéliser ces deux dimensions.

	Clustering hiérarchique		**Clustering spectral**	
	ARI	NMI	ARI	NMI
EGraphClustering	0.26	0.38	0.25	0.37
RObject-Based Clustering	0.23	0.31	0.17	0.28
Pixel-Based Clustering	0.15	0.31	0.22	0.36
Pixel-Object-Based Clustering	0.25	0.43	0.22	0.37

TABLE 4.2 – Les valeurs du NMI et ARI des quatres approches appliquées au site de la Baisse Plaine de l'Aude combinées aux deux algorithmes de clustering : l'algorithme spactral et hiérachique.

	Clustering hiérarchique		**Clustering spectral**	
	ARI	NMI	ARI	NMI
EGraphClustering	0.43	0.53	0.52	0.59
RObject-Based Clustering	0.24	0.32	0.2	0.34
Pixel-Based Clustering	0.35	0.34	0.25	0.38
Pixel-Object-Based Clustering	0.39	0.38	0.32	0.43

TABLE 4.3 – Les valeurs du NMI et ARI des quatres approches appliquées au site de la Vallée du Libron combinées aux deux algorithmes de clustering : l'algorithme spectral et hiérarchique.

Nous avons aussi évalué les performances des différentes approches en terme de temps d'exécution. Les tableaux 4.2 et 4.3 présentent le temps d'exécution des différentes méthodes pour la Basse Plaine de l'Aude et la Vallée du Libron respectivement. Le processus d'analyse des images satellites est scindé en quatre étapes : la construction des graphes, la génération des synopsis, le calcule des distances, le clustering (hiérarchique ou spectral).

La construction des graphes d'évolutions et la génération des synopsis sont propres à la méthodes proposée tandis que le calcul des distances et le clustering sont en commun aux quatre méthodes. Les résultats montrent que le *EGraphClustering*est la moins couteuse en terme de temps d'exécution en comparaison des méthodes basées pixel c'est-à-dire *Pixel-Based Clustering* et *Pixel-Object-Based Clustering*. En effet, le nombre de pixel des images brutes est supérieur au nombre d'objets à analyser, cela explique le différence dans le temps d'exécution.

Le *EGraphClustering*nécessite moins d'une seconde pour le clustering hiérarchique et environ une seconde pour le clustering spectral. Tandis que les approches basées pixel, nécessitent plus de 100 secondes pour le clusering hiérarchique et plus de 29000 secondes pour le clustering spectral sur la Basse Plaine de l'Aude. Concernant la Vallée du Libron, ces méthodes requièrent environ 6 secondes pour le clustering hiérarchique et plus de 1000 seconde pour le clustering spectral. Ces résultats montrent que notre méthode est plus adaptée au traitement des séries temporelle d'images satellites de taille importante car elle permet le passage à l'échelle.

Ces résultats soulignent aussi que la *RObject-Based Clustering* est la plus rapide avec le temps d'exécution le plus bas. Néanmoins, elle est la méthode la moins performante parmi les quatre.

	Construction des graphes	Génération des synopsis	Calcul des distance	Clustering hiérarchique	Clustering spectral
EGraphClustering	2.81	0.1	0.27	0.002	1.41
RObject-Based Clustering	-	-	0.09	0.001	1.36
Pixel-Based Clustering	-	-	11622.50	116.73	29891.18
Pixel-Object-Based Clustering	-	-	21337.87	154.88	29980.71

TABLE 4.4 – Temps d'exécution (en secondes) des différentes approches sur la Basse Plaine de l'Aude.

	Construction des graphes	Génération des synopsis	Calcul des distance	Clustering hiérarchique	Clustering spectral
EGraphClustering	0.36	0.02	0.05	0.003	1.04
RObject-Based Clustering	-	-	0.02	0.0007	0.96
Pixel-Based Clustering	-	-	1738.02	6.48	1185.30
Pixel-Object-Based Clustering	-	-	3013.02	6.04	1153.35

TABLE 4.5 – Temps d'éxecution (en secondes) des différentes approches sur Vallée du Libron.

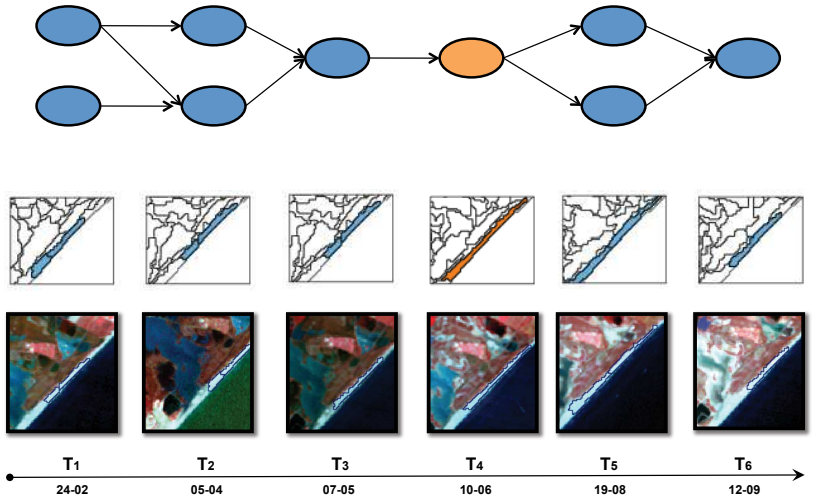
4.3.4 Résultats de l'évaluation qualitative

Afin de mieux caractériser les groupes d'entités spatio-temporelles détectés par notre méthode, nous avons manuellement inspecté les différents clusters résultants. Nous avons sélectionné quelques graphes d'évolution types dans chaque cluster dont nous avons analysé le type d'évolution. Les exemples sont illustrés dans les Figures 4.5-4.8.

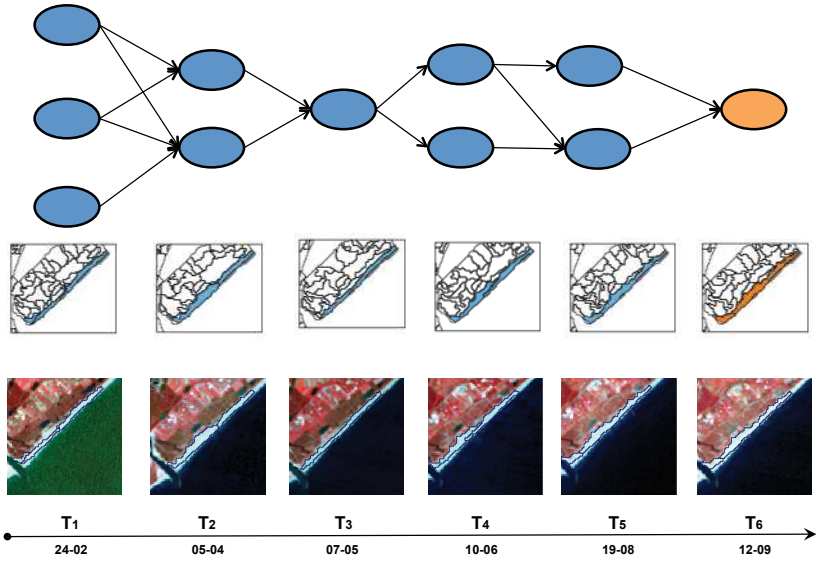
Les Figure 4.5 et 4.6 illustrent des graphes d'évolution de la Basse Plaine de l'Aude. Les graphes d'évolution de la Figure 4.5 appartiennent au même cluster tandis que les deux graphes de la Figure 4.6 appartiennent à un cluster différent. Les deux clusters correspondent aux classe *plages* et *lagune littoral* respectivement. La couleur des objets est reliée à leurs valeurs radiométiques.

Les graphes d'évolution de la Figure 4.5 correspondent à des *plages*. Les *plages* sont caractérisées par une stabilité radiométrique. Elles présentent les mêmes valeurs radiométriques dans toutes les images de la série. En effet, la nature sableuse des plages les rend insensible aux changements saisonnales, ce qui affecte aussi la structure du graphe qui est caractérisé par sa simplicité, c'est-à-dire, structure linéaire. Quant aux graphes d'évolution de la Figure 4.6, ils représentent des *lagunes littorales*, leur évolution est caractérisée par un changement de valeurs radiométriques. Les deux graphes présentent des valeurs radiométriques stables au début de la série (T_1, T_2, T_3), ces derniers varient à la fin de la série (T_4, T_5, T_6). Ce phénomène est dû au changement saisonnier. En période d'hiver, les lagunes sont humides remplies d'eau tandis qu'en été, elles deviennent sèches et contiennent moins d'eau. L'apparition des zones sèches influence aussi la structure du graphe. L'objet représentant

la lagune au début de la série devient fragmenté en plusieurs objets à la fin de la série.

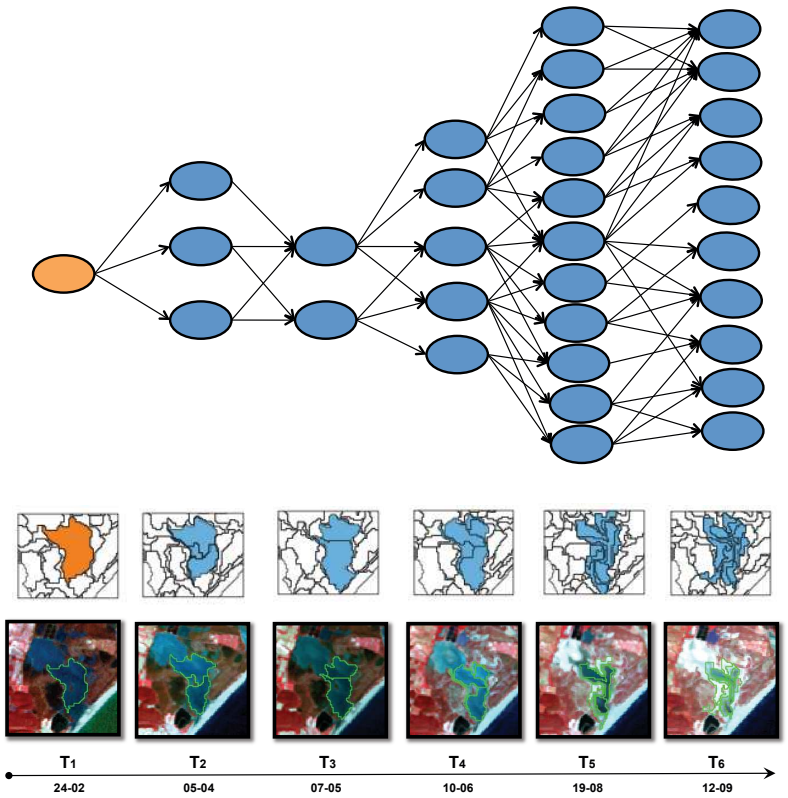


(a)

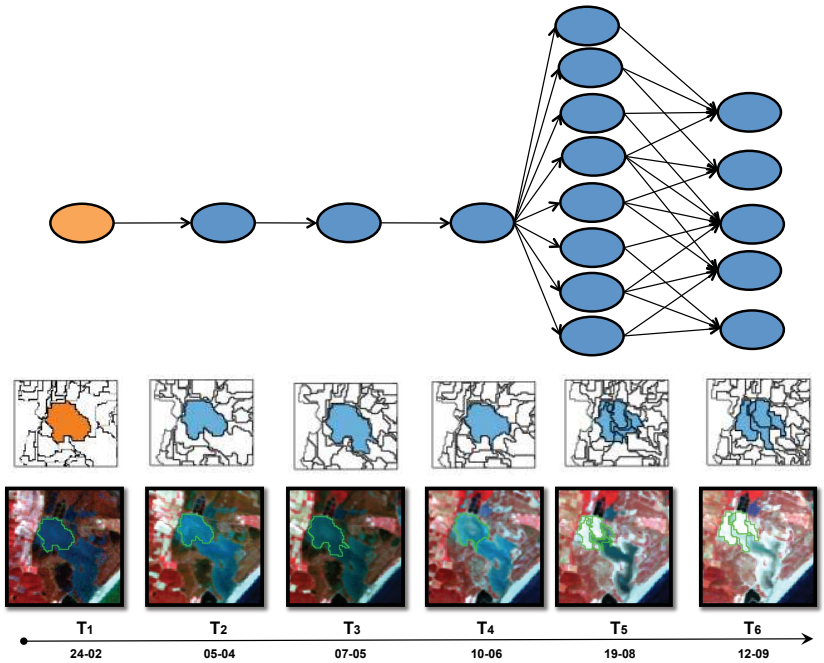


(b)

FIGURE 4.5 – Exemples de graphes d'évolution identifiés sur la Basse Plaine de L'Aude : le graphe (a) représente l'évolution d'une surface de *plage* identifiée dans la quatrième image T_4 et le graphe (b) représentent l'évolution d'une surface de *plage* identifiée dans la dernière image T_6.

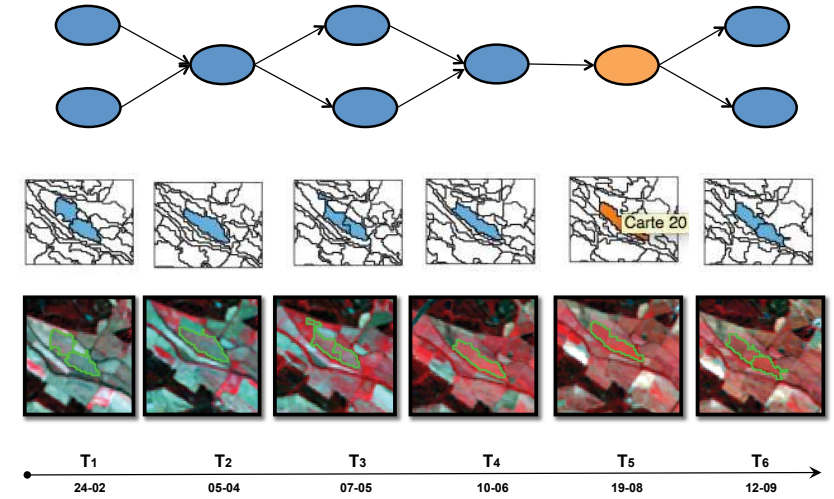


(a)

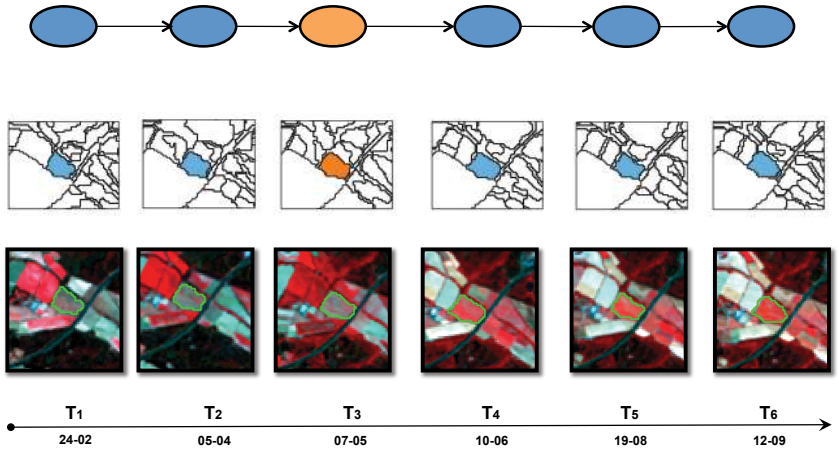


(b)

FIGURE 4.6 – Exemples de graphes d'évolution identifiés sur la Basse Plaine de L'Aude : le graphe (a) et le graphe (b) représentent l'évolution d'une *lagune littoral* identifiée dans la première image de la série T_1.

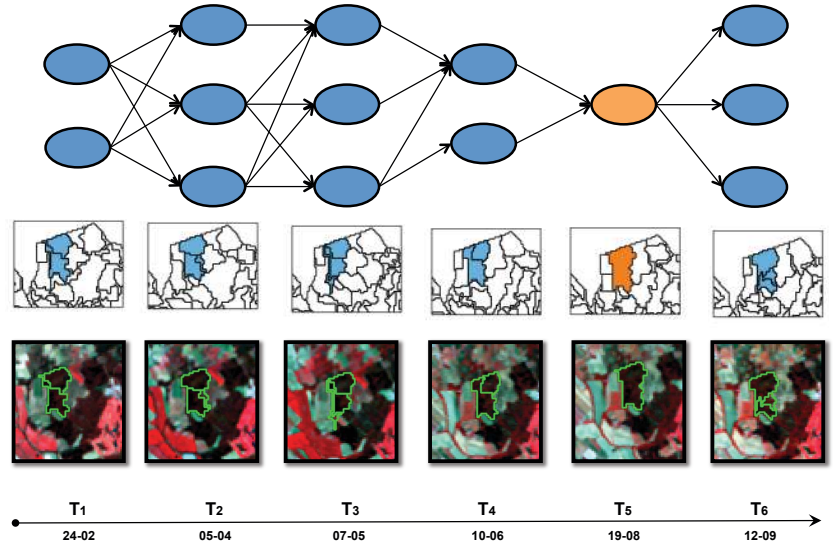


(a)

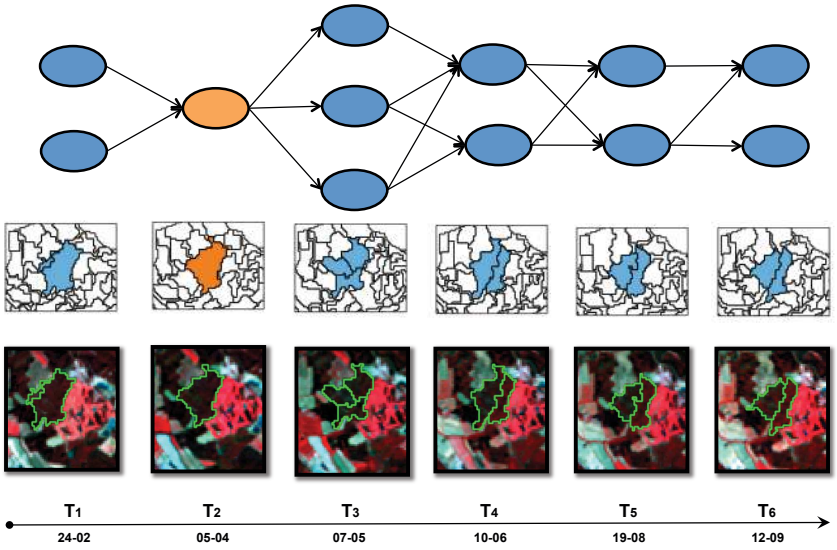


(b)

FIGURE 4.7 – Exemples de graphes d'évolution identifiés sur la Vallée du Libron : le graphe (a) représente l'évolution d'une parcelle de *vignoble* identifiée dans la cinquième image de la série T_5 et le graphe (b) représentent l'évolution d'une parcelle de *vignoble* identifiée dans la troisième image T_3.



(a)



(b)

FIGURE 4.8 – Exemples de graphes d'évolution identifiés sur la Vallée du Libron : le graphe (a) représente l'évolution d'une surface de *végétation chlorophylle* identifiée dans la cinquième image de la série T_5 et le graphe (b) représentent d'une surface de *végétation chlorophylle* identifiée dans la deuxième image T_3.

Les Figures 4.7 et 4.8 illustrent des exemples de la Vallée du Libron. De même les graphes d'évolution de la Figure 4.7 appartiennent au même cluster, celui-ci correspond à la classe *vignoble*. Les graphes d'évolution de la Figure 4.8 appartiennent à un autre cluster qui correspond à la classe *végétation chlorophylle*.

Les graphes d'évolution de la figure 4.7 se caractérisent par une évolution similaire. Leurs valeurs radiométrqiues sont stables au début de la série (T_1, T_2) ainsi qu'a la fin de la série (T_5, T_6). Les deux graphes d'évolution présentent un changement de valeurs radiométriques au milieu de la série (T_3, T_4). Cette dynamique correspond au cycle phonologique de cette culture.

Les graphes d'évolution de la Figure 4.8 montrent une évolution différente. En ce qui concerne la réponse spectrale de la *végétation chlorophylle*, elle demeure stable. En effet, ce type de végétation est caractérisé par un cycle de croissance très long qui s'étale sur des années qui ne peut être illustré à travers une série temporelle annuelle couvrant huit mois.

L'inspection des clusters appuie les résultats de l'évaluation quantitative et atteste de la pertinence de la méthode proposée pour l'identification et la caractérisation les évolutions des entités spatio-temporelles à partir des STIS.

4.4 Conclusion

Ce chapitre décrit la méthode *EGraphClustering*, une nouvelle approche pour l'analyse non supervisée de séries temporelles d'images satellites basée objet. Notre objectif est de mettre en évidence des patrons d'évolution dans les séries temporelles d'images satellites.

Cette approche explore une série temporelle segmentée, elle analyse d'abord les objets de la série afin de sélectionner des entités spatio-temporelles (objets de référence) d'intérêt. L'évolution des ces entités spatio-temporelles est ensuite illustrée en utilisant des graphes, nommés graphes d'évolution. Les graphes d'évolution permettent de décrire les entités spatio-temporelles dans chaque image de la série temporelle. Ils sont transformés en synopsis. Le synopsis est une structure qui permet de résumer l'information radiométrique des graphes d'évolution. Ils permettent d'estimer la distance entre les graphes d'évolution qui leur correspondent et de générer une matrice de distance. Enfin les entités spatio-temporelles sont partitionnées en groupes homogènes en utilisant des algorithmes de clustering sur la matrice de distance. Le clustering vise à identifier et mettre en évidence des évolutions similaires d'entités spatio-temporelles.

Cette méthode a été évaluée sur deux sites d'étude : la Basse Plaine de l'Aude et la Vallée du Libron. Le clustering des graphes d'évolution à été comparé à une classification experte. Les résultats montrent la pertinence des graphes d'évolution pour l'analyse des dynamiques dans les STIS. L'inspection des clusters a permis de caractériser les différents patrons d'évolution qui ont été identifiés. La méthode *EGraphClustering* a été aussi comparée aux méthodes de l'état de l'art basées pixel. Les résultats ont démontré que l'approche d'analyse par objet est compétitive par rapport aux approches par pixel, voire plus performante en terme de temps d'exécution.

Le chapitre suivant sera dédié à l'analyse des séries temporelles pluriannuelles,

nous présentons les différentes problématiques liées à l'information pluriannuelle et comment la méthode présenté ci-dessus a été adaptée pour l'identification des évolutions pluriannuelles.

Chapitre 5

Analyse de séries temporelles d'images satellites pluriannuelles

Sommaire

5.1 Introduction

L'objectif de cette thèse est l'étude de la dynamique des habitats naturels et semi-naturels en exploitant des séries temporelles d'images satellites.

Dans le Chapitre 4, nous avons introduit une méthode d'analyse de séries temporelles d'images satellites basée objet. Cette méthode considère des STIS segmentées et permet d'identifier des entités spatio-temporelles d'intérêt. L'évolution de ces entités spatio-temporelle est représentée par des graphes nommés graphes d'évolution. Ces graphes d'évolution sont analysés afin de mettre en évidence des patrons d'évolution ainsi que des entités spatio-temporelle évoluant similairement. Afin de valider cette méthode, nous avons considéré deux sites d'étude : la Basse Plaine de l'Aude et la Vallée du Libron décrits par des séries temporelles d'images satellites annuelles.

Dans ce chapitre, nous proposons une méthode d'analyse de séries temporelles d'images satellites pluriannuelles. Les STIS pluriannuelles permettent de caractériser des phénomènes récurrents (comme par exemple le cycle de culture) ainsi que des phénomènes avec des cycles d'évolution long. Nous avons adapté la méthode décrite dans le chapitre 4 pour l'analyse de séries temporelles pluriannuelles selon deux points principaux : (i) Nous avons introduit des contraintes dans l'étape de construction des graphes d'évolution afin de considérer uniquement les images pertinentes ne contenant pas de redondance, (ii) Nous avons proposé une méthode de calcul de distance basée sur la mesure Dynamique Time Warpping (Sakoe et Chiba, 1978).

En outre, nous avons défini une approche adaptée à l'analyse multi-site. Nous nous intéressons dans ce chapitre à l'identification des similarités intra-sites et inter-sites c'est-à-dire, nous analysons plusieurs sites pour identifier les entités spatio-temporelles évoluant similairement dans un même site mais aussi dans tous les sites.

La méthode proposée a été évaluée sur trois sites d'étude : La Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en exploitant des séries temporelles d'images satellites acquissent sur une période de dix-huit ans de 1990 jusqu'en 2008.

Dans ce chapitre, nous présentons la méthode proposée pour l'analyse des séries temporelles pluriannuelles et multi-site. Nous décrivons l'étape de détection des entités spatio-temporelles, l'étape de construction des graphes d'évolution et enfin l'étape de clustering (Section 5.2). Nous décrivons aussi l'approche de validation ainsi que les résultats obtenus (Section 5.3).

5.2 Analyse non supervisée des STIS pluriannuelles

Dans cette section, nous présentons le processus global de la méthode proposée pour l'analyse des séries temporelles d'images satellites pluriannuelles combinant plusieurs sites d'étude. Nous décrivons ensuite chacune des différentes étapes en détail.

5.2.1 Vue globale de la méthode

Notre approche d'analyse de séries temporelles d'images satellites pluriannuelles compte trois étapes. La Figure 5.1 illustre ces trois étapes qui sont : (i) La détection des entités spatio-temporelles, (ii) la construction des graphes d'évolution et (iii) le clustering des graphes d'évolution. Pour une analyse inter-sites, plusieurs séries temporelles d'images satellites sont considérées ainsi que leur segmentation (Figure 5.1(Entrée)).

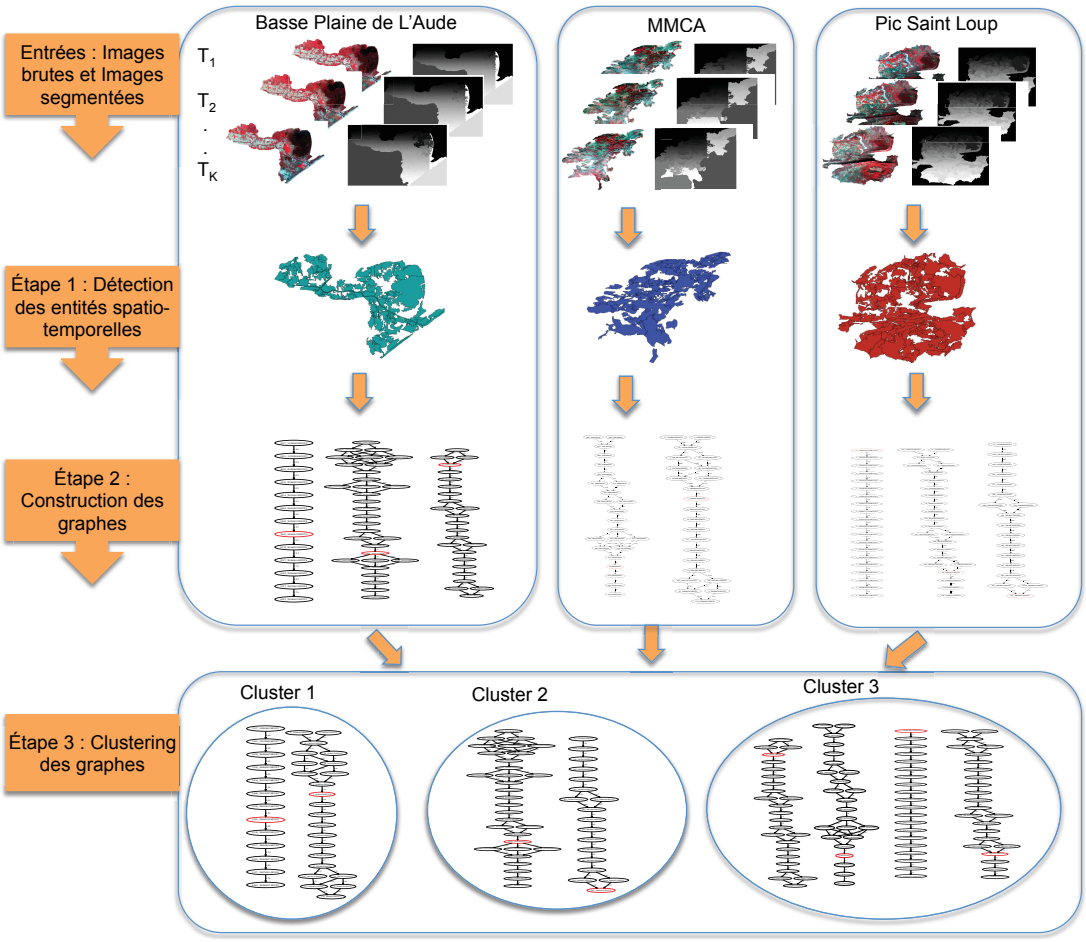


FIGURE 5.1 – Schéma général illustrant les différentes étapes d'analyse des entités spatio-temporelles (objets de référence) pour une étude multi-site.

La première étape est la détection des objets de référence/entités spatio-temporelles (Figure 5.1(Étape 1)). Pour chaque objet de référence détecté précédemment, un graphe d'évolution est construit (Figure 5.1(Étape 2)). Les graphes d'évolution permettent d'illustrer l'évolution des objets de référence. La dernière étapes est le clustering des graphes (Figure 5.1(Étape 3)), les graphes d'évolution sont analysés afin d'identifier des groupes homogènes qui évoluent similairement.

L'analyse multi-site implique plusieurs sites d'étude, lors des deux premier étapes chaque site est analysé indépendamment. Pour chaque site, on extrait les objets de référence puis on construit les graphes d'évolution. Lors de la dernière étape, les graphes d'évolution construits sur tous les sites sont rassemblés puis partitionnés en groupes homogènes pour identifier et mettre en évidence des similarités inter-sites.

5.2.2 Détection des entités spatio-temporelles

Ainsi que défini dans la Section 4.2.2, les entités spatio-temporelles correspondent aux objets de référence. Ils représentent les objets d'intérêt à analyser. Il peut s'agir d'une parcelle agricole, une zone forestière, un lac etc. Lors de la détection des objets de référence, tous les objets des images satellites sont considérés. Les objets les plus pertinents sont sélectionnés itérativement de manière à couvrir la zone d'étude au mieux en minimisant le chevauchement. À chaque itération, les objets sont pondérés, l'objet dont le poids est maximal est sélectionné comme un nouvel objet de référence. Afin de minimiser le chevauchement entre les objets de référence, un paramètre α est utilisé. Ainsi les objets dont le poids est inférieur à α sont écartés. La détection des objets de références est décrit par l'Algorithme 5.

L'approche proposée se situe dans un contexte multi-site où plusieurs sites sont analysés afin d'identifier et de mettre en évidence des évolutions pluriannuelles similaires entre leurs entités spatio-temporelles. L'étape de détection de ces entités spatio-temporelles analyse chaque site d'étude indépendamment. Pour chaque site, on inspecte tout l'ensemble d'objets afin d'en sélectionner un sous-ensemble d'objets de référence. L'évolution de ces objets de référence est par la suite illustrée par un graphe d'évolution.

5.2.3 Construction des graphes d'évolution

Une fois les objets de référence de chaque sites sélectionnés, pour chaque objet de référence, nous construisons un graphe d'évolution qui illustre son évolution dans la série temporelle. Ainsi que défini en Section 4.2.3, un graphe d'évolution est un graphe acyclique orienté défini par, $G_{o^*} = (V_{o^*}, E_{o^*})$. L'ensemble V_{o^*} représente les nœuds de ce graphe et l'ensemble E_{o^*} représente ses arcs.

La construction des graphes d'évolution compte trois étapes : la sélection des nœuds, l'organisation des nœuds en couche et la définition de l'ensemble d'arcs entre les nœuds.

La sélection des nœuds se base sur l'étendue spatiale de l'objet de référence (Les pixels couverts par l'objet de référence). Les objets de la série temporelle sont analysés afin d'identifier ceux qui chevauchent l'objet de référence. Deux seuils de chevauchement sont définis afin de sélectionner uniquement les objets ayant un taux de chevauchement élevé avec l'objet de référence. L'ensemble des nœuds du graphe G_{o^*} est défini par l'Équation 4.1 (Section 4.2.3). Dans les travaux proposés par (Guttler et collab., 2017), les auteurs exploitent la série temporelle complète pour

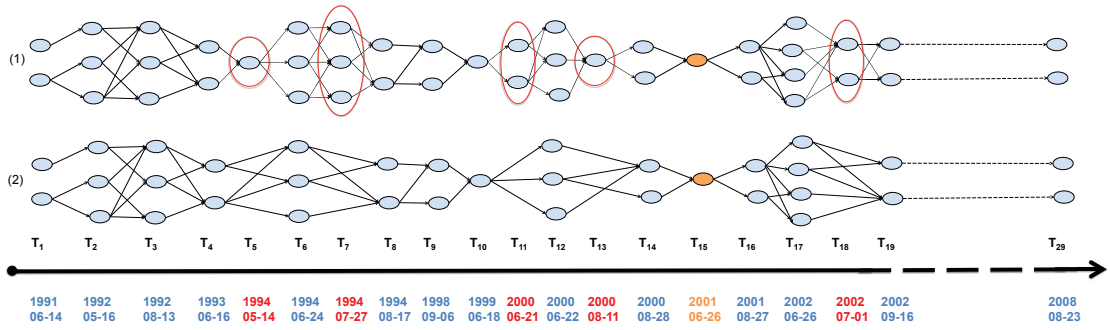


FIGURE 5.2 – Exemple de deux graphes d'évolution de la Basse Plaine de l'Aude illustrant l'évolution d'un objet de référence sélectionné sur l'image acquise en 26-06-2011. Le graphe (1) exploite toutes les images de la série tandis que le graphe (2) exploite un sous ensemble uniquement.

l'identification des nœuds du graphe d'évolution. En effet, la série temporelle considérée dans cet article est annuelle, elle compte six images acquises sur une période de huit mois dans l'objective d'analyser des phénomènes phénologiques. À la différence de ces travaux, notre analyse porte sur des sites décrits par des séries temporelles pluriannuelles. L'exploitation de toutes les images de la série peut introduire de la redondance d'information. En effet, la période entre le temps d'acquisition des images peut être négligeable. Par example les dates d'acquisition de ces trois images satellites de la MMCA : 21-08-2001, 22-08-2001 and 27-08-2001, indiquent qu'elles sont acquises sur une période d'une semaine.

Afin d'éviter la redondance de l'information portée par les graphes d'évolution, nous avons introduit une contrainte sur la période entre l'acquisition de deux images satellites successives. Deux images trop similaires (dont les dates d'acquisition sont trop proches) ne sont pas considérées toutes les deux dans la phase de construction des graphes d'évolution. En s'appuyant sur les recommandations des experts, nous avons fixé la période minimale entre les dates d'acquisition des images satellites à deux mois. En se basant sur la date d'acquisition de l'image de l'objet de référence, on considère uniquement les images satellites successives ayant une différence entre leurs dates d'acquisition d'au moins deux mois.

Une fois les objets du graphe sélectionnés, ils sont d'abord organisés en groupes en fonction de la date d'acquisition de leur image satellite. Ces groupes sont ensuite ordonnés par ordre croisant. Des arcs sont enfin définis entre les objets de chaque paire de groupe successive s'ils se chevauchent. L'ensemble des arcs du graphe G_{o^*} est défini par l'équation 4.2 (Section 4.2.3).

La Figure 5.2 illustre deux exemples de graphes d'évolution (1,2) correspondant au même objet de référence. Le graphe d'évolution (1) est construit en utilisant toute la série temporelle, tandis que le graphe (2) est construit uniquement en utilisant les images satisfaisant la contrainte sur la différence entre le temps d'acquisition des images qui est fixé à deux mois.

Les objets d'un graphe d'évolution sont organisés de gauche à droite en se basant sur les dates d'acquisition de leur image satellite. L'objet orange correspond à l'objet de référence, tandis que les objets bleu correspondent aux objets qui chevauchent l'objet de référence. La date d'acquisition en orange correspond à celle de l'image de l'objet de référence. Les dates d'acquisition en bleu correspondent aux images satellites considérées lors de la construction du graphe d'évolution. Tandis que les dates d'acquisition en rouge correspondent aux images écartées lors de la construction du graphe d'évolution car elles ne satisfont pas la condition sur la période entre les dates d'acquisition. Le graphe (1) met en évidence les objets écartés (encerclé en rouge) qui n'apparaissent pas dans le graphe d'évolution final (2). Les objets des graphes d'évolution illustrés dans la Figure 5.2 sont reliés par des arcs, seul les objets dont les images se succèdent sont reliés.

Enfin, les graphes d'évolution construit sur tous les sites d'étude sont combinés lors du clustering afin d'identifier des similarités inter-sites.

5.2.4 Clustering des graphes d'évolution

Le clustering des graphes est la dernière étape de la méthode d'analyse multi-site. De même que dans le Chapitre 4, les graphes d'évolution sont d'abord transformés en synopsis. Les synopsis permettent de calculer la distance entre les graphes d'évolution. Ils résument l'information radiométrique portée par les objets des graphes d'évolution.

Les synopsis correspondent à une séquence d'objets $\widetilde{O_{T_l}}$, les objets du synopsis sont calculés à partir des objets du graphe d'évolution lui correspondant. Le calcul des objets $\widetilde{O_{T_l}}$ est défini par l'équation 5.1.

$$Info(\widetilde{O_{T_l}}) = \frac{\sum_{o_{T_l} \in E_{o^*}} |Pix(o_{T_l})| \; Info(o_{T_l})}{\sum_{o_{T_l} \in E_{o^*}} |Pix(o_{T_l})|} \tag{5.1}$$

Nous avons introduit une nouvelle formule pour le calcul de synopsis, elle permet de les générer dans un temps raisonnable. En effet, la formule de construction des synopsis introduite dans le Chapitre 4 se base sur l'extraction des chemins. L'appliquer aux graphes d'évolution pluriannuelles de taille importante implique un coût élevé en terme de temps d'exécution.

Les graphes d'évolution sont structurés sous forme de couches. Les objets de chaque couche sont agrégés, leurs valeurs radiométriques sont pondérées par leur taille puis moyennées. Sachant que les objets de chaque couche appartiennent à une seule image, le synopsis contiendra autant d'objets que d'images satellites de la série temporelle. Chaque objet $\widetilde{O_{T_l}}$ décrit l'objet de référence dans l'image I_{T_l}.

Le processus d'extraction des synopsis est illustré dans la Figure 5.3. Ce schéma illustre un graphe d'évolution et le synopsis qui lui correspond. Ce graphe d'évolution se compose de quatre couches d'objets : $\{\{o^1_{T_1}, o^2_{T_1}\}, \{o^1_{T_2}, o^2_{T_2}\}, \{o^1_{T_3}\}, \{o^1_{T_4}, o^2_{T_4}\}\}$. Les objets de chaque couche sont agrégés en un seul objet : $\{\widetilde{O_{T_1}}, \widetilde{O_{T_2}}, \widetilde{O_{T_3}}, \widetilde{O_{T_4}}\}$

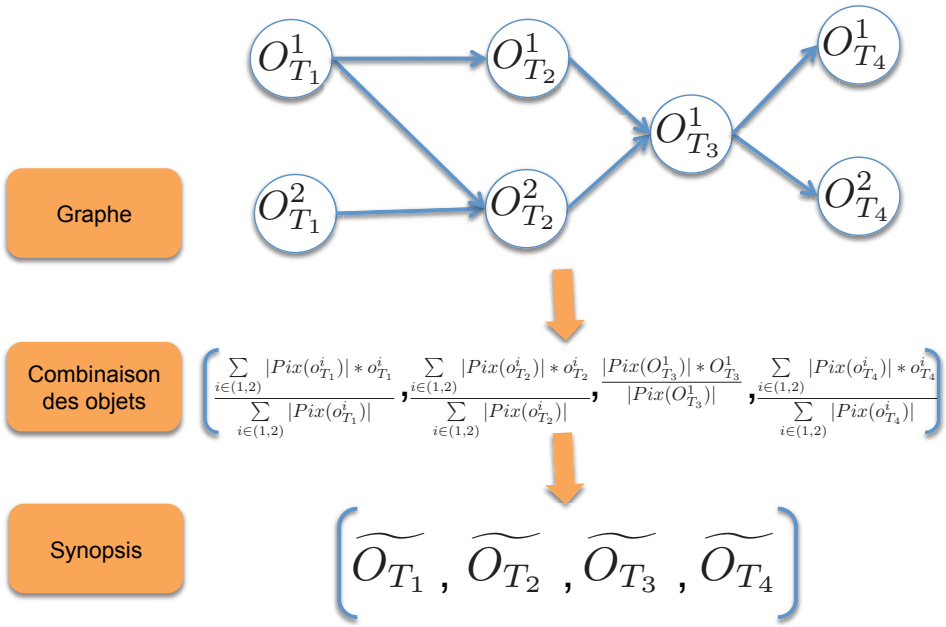


FIGURE 5.3 – Procédure de génération des synopsis à partir des graphes d'évolution.

Chaque graphe d'évolution est transformé en synopsis. Pour chaque paire de synopsis, on calcule la distance entre leurs objets en se basant sur les valeurs radiométriques de ces derniers. Le calcul de distance se base sur la mesure Dynamic Time Warping (DTW) (SAKOE et CHIBA, 1978). Étant donnés deux graphes d'évolution G_1 et G_2 et leur synopsis respectifs syn_1 et syn_2, leur distance est calculée selon l'Équation 5.2.

$$dist_s(syn_1, syn_2) = DTW(syn_1, syn_2) \tag{5.2}$$

La mesure Dynamic Time Warpping permet de déterminer l'alignement optimal entre deux séries temporelles de taille différente. Ainsi, elle est adaptée au traitement de séries temporelles de longueur différente.

Nos avons utilisé cette mesure étant donné que les séries temporelles d'images satellites acquises sur les trois sites d'étude : Basse Plaine de L'Aude, la MMCA et le PIc Saint Loup, sont caractérisées par un échantillonnage irrégulier (dates d'acquisitions différentes) par année et par mois comme illustré dans les Figures 3.9-3.11. Cette mesure permet de palier ce problème.

Comme exemple, nous considérons deux séries temporelles $C = (c_1, c_2, ..., c_n)$ et $Q = (q_1, q_2, ..., q_m)$ de longueur n et m respectivement. Afin d'estimer la distance entre C et Q, DTW se base sur une fonction de distance locale qui calcule la distance entre chaque paire d'éléments des deux séries. Une matrice de coût de taille $n * m$ est générée comme défini par l'équation 5.3.

$$D(c_l, q_r) = \theta(c_l, q_r) + min \begin{cases} D(c_{l-1}, q_{r-1}) \\ D(c_l, q_{r-1}) \\ D(c_{l-1}, q_r) \end{cases} \tag{5.3}$$

Où :

— θ représente la fonction de distance locale.

Le coût de l'alignement optimal est donné par le dernier élément de la matrice, cette valeur représente la distance entre les deux séries temporelles (Équation 5.4).

$$DTW(C, Q) = D(c_n, q_m) \tag{5.4}$$

La matrice de coût permet aussi d'identifier le chemin de déformation. Celui-ci représente les différent coûts optimaux à partir de l'élément $(D(c_0, q_0))$ jusqu'à l'élément $D(c_n, q_m)$ comme illustré au travers de la Figure 5.4.

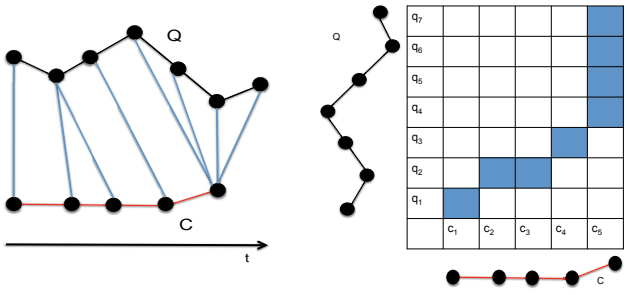


FIGURE 5.4 – Exemple de la matrice de coût calculée pour les deux séquences C and Q et illustration du chemin de déformation identifié.

Cette mesure permet ainsi de détecter le déphasage entre les deux séries et de les aligner. Elle a été utilisée initialement dans le domaine de la reconnaissance vocale et du traitement du signal (MÜLLER, 2007).

En se basant sur la mesure DTW, nous avons calculé une matrice de distance qui compte la distance entre chaque paire de synopsis. Cette matrice est exploitée pour le clustering des graphes d'évolution. Lors de cette phase, nous combinons les graphes d'évolution construits sur les trois sites d'étude (Basse Plaine de L'Aude, la MMCA et Pic Saint Loup) afin de réaliser une analyse inter-sites. Le clustering permet d'identifier des groupes de graphes d'évolution dont les dynamiques sont similaires ainsi que des similarités inter-sites. Il existe dans la littérature plusieurs algorithmes de clustering (clustering kmeans, clustering hiérarchique, clustering spectral, etc) qui permettent d'analyser les graphes d'évolution en se basant uniquement sur la matrice de distance construite, indépendamment des données initiales (Image satellites).

5.3 Expérimentations

Notre objectif est de réaliser une analyse inter-sites, nous avons présenté dans les sections précédentes une approche qui permet d'analyser plusieurs sites simultanément afin d'identifier des similarités inter-sites. Afin de valider notre approche, nous avons mené nos expérimentations sur trois sites d'étude qui sont : Basse Plaine de L'Aude, la MMCA et le PIc Saint Loup en se basant sur les séries temporelles d'images satellites multi-annuelles Spot, décrites dans la Section 3.3.2. Dans cette section, nous présentons d'abord l'approche de sélection des paramètres. Nous reportons ensuite les résultats de l'analyse des graphes d'évolution. Finalement, nous décrivons quelques graphes d'évolution similaires identifiés sur les trois sites d'étude.

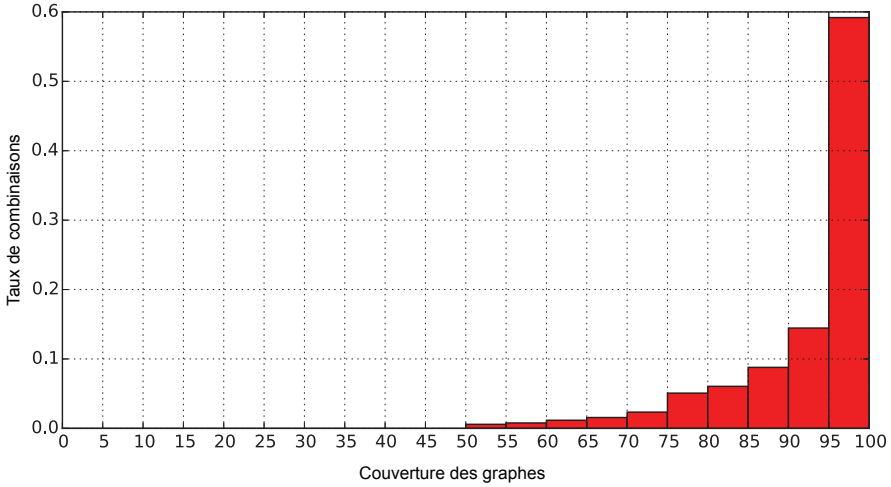
5.3.1 Sélection des paramètres

L'approche d'analyse des graphes d'évolution se base sur trois paramètres. Le paramètre α qui intervient lors de la sélection des objets de référence. Les deux paramètres $sigma_1$ et $sigma_2$ qui interviennent dans l'étape de construction des graphes d'évolution. Comme décrit dans la Section 4.3.1, la sélection des valeurs de ces trois paramètres se base sur la couverture ainsi que le chevauchement des graphes d'évolution. De même, nous avons fait varier les trois paramètres dans l'intervalle $[0, 1]$ avec un pas de 0.1. Pour chaque combinaison de paramètres, nous avons rapporté la couverture ainsi que l'intersection des graphe d'évolution. Nous avons enfin sélectionné des seuils de couverture pour chaque zone d'étude. La Figure 5.5 illustre le taux de couverture des graphes d'évolution des différentes combinaisons sur la Basse Plaine de L'Aude (a), la MMCA (b) et le Pic Saint Loup (c).

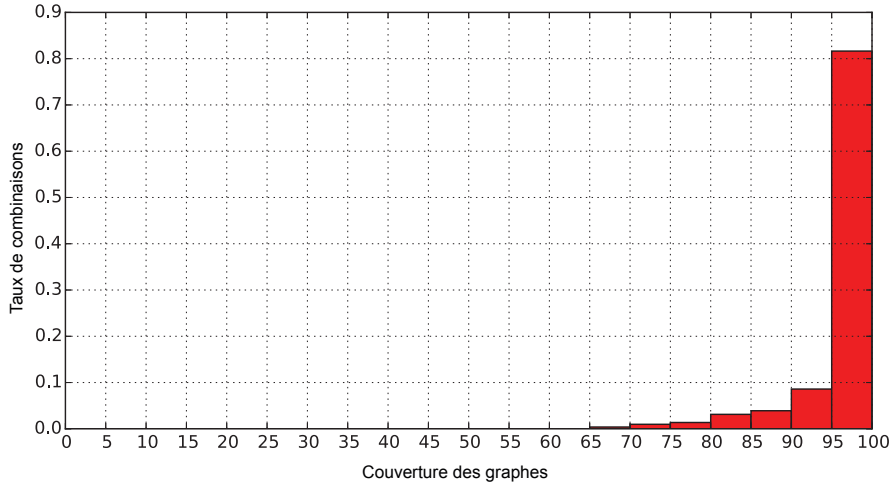
En se basant sur les résultats obtenus, plus de 50% des combinaisons présentent un taux de couverture de plus de 95% de la zone d'étude pour les trois sites. Cependant, ces combinaisons présentent aussi un taux de chevauchement élevé. Afin de minimiser le taux de chevauchement, les seuils sélectionnées sont de 85%, 90% et 90% sur la Basse Plaine de L'Aude, la MMCA et le Pic Saint Loup respectivement. Les seuils définis ont permis de sélectionner les valeurs des trois paramètres. Ces valeurs sont rapportées dans le tableau5.1. Ce tableau définit aussi le taux de couverture et d'intersection des graphes d'évolution résultant pour chaque site d'étude.

	α	σ_1	σ_2	Graphes d'évolution	Couverture des graphes d'évolution	Chevauchement des graphes d'évolution
Aude Valley	0.5	0.9	0.7	67	88.27	26.02
MMCA	0.4	0.9	0.8	79	90.29	22.71
Pic Saint Loup	0.3	0.9	0.8	87	90.80	30.3

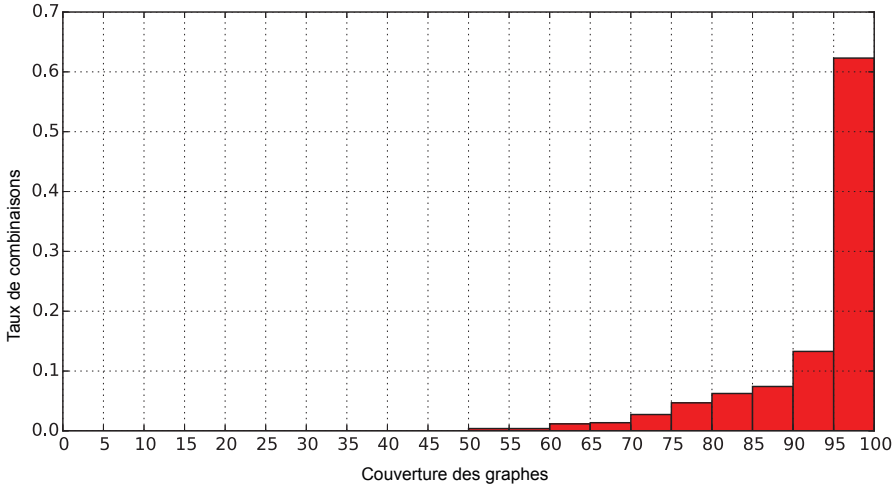
TABLE 5.1 – Les valeurs des paramètres α, σ_1 et σ_2 définis sur les trois sites d'étude. Ainsi que le nombre de graphes d'évolution construits, leur taux de couverture et leur taux de chevauchement.



(a)



(b)



(c)

FIGURE 5.5 – Taux de couverture des graphes d'évolution générés par les différentes combinaisons sur les trois sites d'étude : (a) La Basse Plaine de l'Aude, (b) la MMCA et (c) le Pic Saint Loup.

Les valeurs des paramètres ont permis d'illustrer l'évolution des zones d'étude au travers de 233 graphes d'évolutions : 67 graphes sur la Basse Plaine de L'Aude, 79 graphes sur la MMCA et 87 graphes sur le Pic Saint Loup. Les 233 graphe d'évolution couvrent 89% des trois zone d'études.

5.3.2 Résultats du clustering

Les graphes d'évolution ont été analysés dans le but d'identifier des dynamiques similaires dans les trois sites d'études. L'algorithme de clustering hiérarchique a été utilisé afin de regrouper les graphes d'évolution. Cet algorithme permet de regrouper les différents graphes d'évolution successivement jusqu'à obtenir un seul cluster contenant tout l'ensemble des entités. Il génère une hiérarchie de partitionnement avec différent nombre de clusters. Nous avons retenu différentes hiérarchies de partitionnement avec un nombre de clusters variant entre 9 et 20. Les différentes solutions ont été évaluées par un expert. Dans ce qui suit nous présentons le partitionnement à 20 clusters qui a été sélectionné étant le plus représentatif des évolutions des trois zones d'étude. Le clustering des graphes d'évolution est basé sur leurs attributs, c'est-à-dire les valeurs radiométriques de leurs objets. Différents descripteurs radiométriques ont été exploités pour la description des objets. Nous avons évalué la pertinence de l'information radiométrique pour la caractérisation des types d'évolutions des entités spatio-temporelles. Dans un premiers cas, les trois bandes spectrales des images Spot : rouge, vert et proche infrarouge, ont été utilisées pour caractériser les objets des graphes d'évolution. Dans un deuxième cas, les objets ont été décrits en utilisant les bandes spectrales combinées aux indices radiométriques : rouge, vert, proche infrarouge, NDVI, NDWI, BI, CI et SAVI. Les Figures 5.6 et 5.7 présentent les résultats du clustering basé sur les deux modalités de description bandes spectrales ainsi que bandes spectrales combinées aux indices radiométriques.

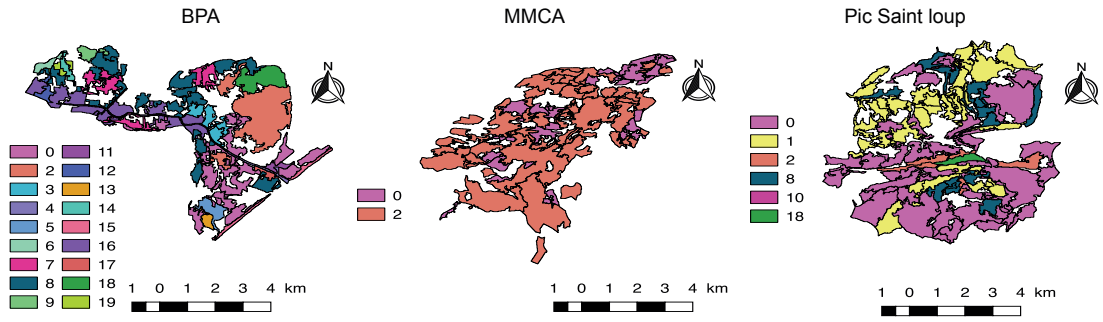


FIGURE 5.6 – La répartition des entités spatio-temporelles dans les clusters résultats (20 clusters) dans les trois sites d'étude en utilisant les bandes spectrales : rouge, verte et infrarouge.

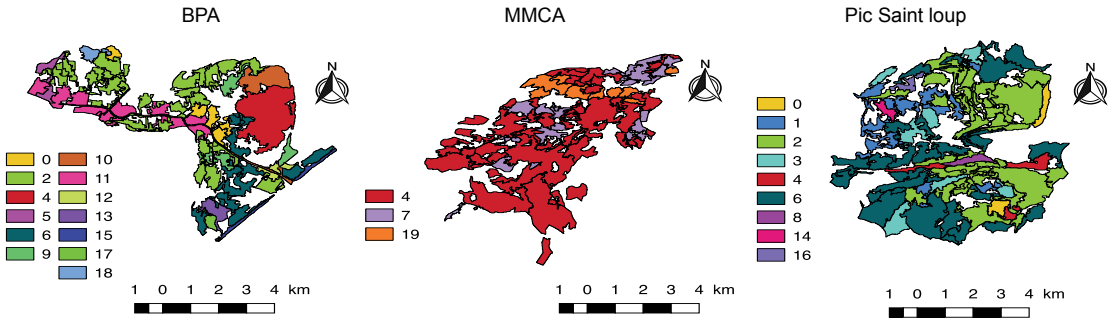


FIGURE 5.7 – La répartition des entités spatio-temporelles dans les clusters résultats (20 clusters) dans les trois sites d'étude en utilisant les bandes spectrales combinées aux indices radiométriques : rouge, verte, proche infrarouge, NDVI, NDWI, BI, CI et SAVI .

A) **La Basse Plaine de L'Aude**

Dans le premier cas, les entités spatio-temporelles de la Basse Plaine de L'Aude ont été regroupées en dix huit clusters (Figure 5.6). Les entités de la classe *surface en eau* ont été regroupées dans les clusters 4,5 et 18. Les clusters 11 et 13 rassemblent les entités de la classe *zone humide*. La classe *plage* correspond au cluster 15. Les entités spatio-temporelles de la classe *vignoble* sont réparties dans les clusters 6, 12 et 16. Les clusters 0, 7 et 8 regroupent les entités spatio-temporelles des différentes classes de couverts végétaux (*vignoble*, *végétation arbustive*, et *culture*) ainsi que de la classe *zone humide* et *surface en eau*. Le cluster 2 correspond aux entités spatio-temporelle des classes *zone humide* et *surface en eau* couvertes par de la végétation aquatique. Les entités spatio-temporelles de la classe *culture* sont associées aux clusters 3, 14 et 17.
Dans la deuxième configuration, les entités spatio-temporelles de la Basse Plaine de l'Aude sont réparties en treize clusters (Figure 5.7). Les indices spectraux ont permis de discriminer les différentes classes des entités spatio-temporelles. Les clusters 10, 13 et 4 correspondent à des entités couvertes par de la végétation aquatique. Les entités de la classe *plage* sont regroupées dans le cluster 15. Les clusters 9 et 17 regroupent les entités de la classe *zone humide*. Les entités des classes de végétation (*culture* et *vignoble*) sont regroupées dans le cluster 0. Les entités des clusters 2, 6 et 12 correspondent aux classes *culture*, *vignoble*, *surface en eau* et *végétation arbustive* dont les dynamiques sont similaires.

B) **La Montagne de la la Maure et Cause d'Aumelas**

La caractérisation des entités spatio-temporelles avec les bandes spectrales a permis de regrouper les entités spatio-temporelles de la MMCA en deux clusters (Figure 5.6). Le cluster 0 regroupe les entités correspondants aux différentes classes de couverts végétaux (*culture*, *végétation arbustive*, *vignoble* et *espace ouvert avec peu ou sans végétation*). L'évolution de ces entités spatio-

temporelles est caractérisée par des valeurs élevées de réflectance dans les trois bandes spectrales (rouge, verte, proche infrarouge). Quant au cluster 2 , il correspond aux entités spatio-temporelles des classes *forêt* et *végétation arbustive* caractérisées par une faible réflectance dans le spectre visible (bandes rouge et verte). Tandis que dans la bande proche infrarouge, ces entités sont caractérisées par une réflectance élevée.
En considérant les deux types d'information, bandes spectrales et indices radiométriques, les entités spatio-temporelles de la MMCA ont été réparties en trois clusters (Figure 5.7). Le NDVI et le SAVI ont permis de discriminer les différentes classes selon la densité de leur couvert végétal. Le cluster 19 est spécifique aux entités spatio-temporelles de la classe *forêt*. Les entités spatio-temporelles des classes forêt et végétation arbustive correspondent au cluster 4. Le cluster 7 regroupe les entités des classes *vignoble* et *végétation arbustive*.

C) **Le Pic Saint Loup**

Concernant le Pic Saint Loup, les entités spatio-temporelles caractérisées par les bandes spectrales ont été réparties en 6 clusters (Figure 5.6). Les deux clusters 2 et 8 correspondent aux entités spatio-temporelles de la classe *forêt*. Le cluster 18 lui aussi correspond aux entités de la classe *forêt*, cependant leurs valeurs radiométriques sont affectées par les zones d'ombre du Pic Saint Loup. Le cluster 1 regroupe les entités spatio-temporelles des différentes classes de couverts végétaux (*culture*, *végétation arbustive* et *vignoble*). Le cluster 2 quant à lui correspond aux entités des classes *forêt* et *végétation arbustive*, qui sont caractérisées par une forte réflectance dans la bande proche infrarouge.
La caractérisation des entités spatio-temporelles du Pic Saint Loup en utilisant les bandes spectrales et les indices spectraux a permis de définir neuf clusters (Figure 5.7). Les clusters 0, 4 et 8 correspondent à des entités spatio-temporelles de la classe *forêt*. Tandis que les clusters 2, 3 et 6 regroupent les entités spatio-temporelles des deux classes *forêt* et *végétation arbustive* qui présentent des évolutions similaires. Les entités spatio-temporelles des classes *culture* et *vignoble* sont associées aux deux clusters 1 et 14.

L'objective de notre étude est l'identification des similarités inter-sites c'est-à-dire, des évolutions spatio-temporelles en commun entre plusieurs sites d'étude. La méthode proposée exploite des séries temporelles d'images satellites pluriannuelles afin d'identifier des entités spatio-temporelles. Les évolutions de ces entités spatio-temporelles sont analysées afin d'identifier des patrons d'évolution inter-sites. Cette méthode est validée sur trois sites d'étude : la Basse Plaine de l'Aude, La MMCA et le Pic Saint Loup. Les résultats nous ont permis de distinguer les dynamiques propres de chaque site d'étude ainsi que les dynamiques similaires entre ces trois sites. Dans ce qui a précédé, nous avons décrit les groupes d'évolution des entités spatio-temporelles identifiées sur chaque site. Dans ce qui suit, nous nous intéressons aux groupes d'évolutions comprenant des entités spatio-temporelles des trois sites.

L'analyse basée sur les bandes spectrales uniquement (Figure 5.6), a permis

d'identifier des similarités entre les trois sites d'étude à travers les deux clusters 0 et 2. Le cluster 0 regroupe des entités spatio-temporelles des différentes classes de végétation. Les entités spatio-temporelles regroupées dans le cluster 0 correspondent aux classes de végétation sur la MMCA et le Pic Saint Loup. Tandis que sur la Basse Plaine de l'Aude, elles correspondent aux classes de végétation ainsi qu'à la classe *zone humide*. Les profils radiométriques de ces entités spatio-temporelles présentent des évolution similaires ce qui explique leur regroupement. La méthode a aussi identifié des similarités entre le Pic Saint Loup et la Basse Plaine de l'Aude, les clusters 8 et 18 regroupent des entités spatio-temporelles des deux sites. Le cluster 18 regroupent des entités spatio-temporelles de la classe *surface en eau* sur la Basse Plaine de l'Aude et à des zone d'ombre sur le Pic Saint Loup. Ces entités spatio-temporelles sont caractérisées par des réponses spectrales similaires sur les bandes rouge et verte Le cluster 8 regroupe principalement des entités spatio-temporelles des classes de végétation sur les deux sites, elles correspondent aux classes *vignoble*, *culture* et *zone humide* sur la Basse Plaine de l'Aude ainsi qu'aux classe *végétation arbustive*, *forêt* et *culture* sur le Pic Saint Loup.

L'analyse des entités spatio-temporelles en utilisant les bandes spectrales combinées aux indices radiométriques (Figure 5.7), a permis de regrouper les entités spatio-temporelles similaires dans le cluster 4. Ces entités correspondent à la classe *forêt* sur le Pic Saint Loup, aux classes *forêt* et *végétation arbustive* sur la MMCA et à de la végétation aquatique sur la Basse Plaine de l'Aude. Elle sont caractérisées par des profils NDVI similaires. Les clusters 2 et 6 regroupent des entités spatio-temporelles de la Basse Plaine de l'Aude et du Pic Saint Loup. Le cluster 2 correspond à des entités des classe *culture*, *vignoble* et *zone humide* sur le premier site ainsi qu'à la classe *forêt* sur le deuxième site. Ces entités spatio-temporelles sont regroupées ensemble car elles présentent des profils NDVI similaires. Le cluster 6 regroupe des entités spatio-temporelles de la classe *végétation arbustive* sur le Pic Saint Loup et la classe *zone humide* sur la Basse Plaine de l'Aude. Ces entités sont caractérisées par une même évolution considérant les trois bandes spectrales mais aussi les indices radiométriques.

Afin d'évaluer les résultats du clustering obtenus et de quantifier les performances de la méthode proposée, nous avons calculé des indices de validité dont l'information mutuelle normalisée et l'indice de rand ajusté (Section 2.2.4.2). Ces indices permettent de comparer les résultats du clustering à la classification experte (Section 3.5.2). Nous avons d'abord comparé les résultats de clustering des entités spatio-temporelles en utilisant les bandes spectrales uniquement à la classification experte. Puis, nous avons comparé les résultats du clustering des entités spatio-temporelles en utilisant les bandes spectrales et les indice radiométriques à la même classification experte. Nous avons considéré plusieurs résultats de partitionnement comptant différents nombres de clusters. Les résultats obtenus sont illustrés dans la Table 5.2

Nous notons que les performance du clustering basé sur les bandes spectrales combinées aux indices radiométriques sont plus élevées que les performance du clus-

	Bandes spectrales		**Bandes spectrales et indices radiométriques**	
	ARI	NMI	ARI	NMI
9 clusters	0.13	0.13	**0.20**	**0.23**
10 clusters	0.13	0.13	**0.17**	**0.26**
15 clusters	0.09	0.17	**0.17**	**0.24**
20 cluster	**0.15**	0.23	**0.15**	**0.24**

TABLE 5.2 – Les résultats de NMI et ARI obtenus en utilisant : (i) les bandes spectrales et (ii) les bandes spectrales combinées aux indices radiométriques.

tering basé sur les bandes spectrales uniquement. En effet, les valeurs NMI et ARI calculées pour le clustering basé sur les bandes spectrales combinées aux indices radiométriques sont plus élevées que celles calculées pour le clustering basé uniquement sur les bandes spectrales. Ces résultats démontrent la pertinence des indices radiométriques pour la caractérisation des évolutions des entités spatio-temporelles. La combinaison des bandes spectrales et des valeurs radiométriques a permis une meilleur discrimination des différents types d'évolution identifiés sur les trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup.

5.3.3 Analyse des graphes d'évolution

Nous avons procédé à l'analyse des graphes d'évolution identifiés par notre méthode sur nos sites d'étude. Les Figures 5.8-5.12 illustrent des exemple de graphes d'évolution. Ces graphes décrivent des exemples de patrons d'évolution identifiés au travers des clusters 4 et 6.

Les graphes d 'évolution identifiés sur les trois sites d'étude se composent de plus de vingt groupes d'objets correspondant aux 20 images de la série temporelle. Afin de rendre les figures lisibles et compréhensibles, nous avons repris cinq images seulement sur la série d'images pour chaque graphe d'évolution et nous avons remplacé les arcs par des pointillés. Nous avons représenté pour chaque graphe d'évolution un ensemble de nœuds. Pour chaque groupe de nœuds, nous avons identifié les objets qui leur correspondent sur les images satellites. La couleur des objets représente les valeurs radiométriques de celui-ci. Le changement de couleur correspond à une évolution de la couverture du sol.

Les Figures 5.8, 5.9 et 5.10 illustrent trois graphes d'évolution de la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup respectivement. Ces trois graphes d'évolution sont regroupés dans le cluster 4. Les deux graphes en Figure 5.9 et 5.10 correspondent à des entités spatio-temporelles de la classe *forêt* caractérisée par un couvert végétal dense. Ainsi les objets leur correspondant dans les images satellites sont caractérisés par une couleur rouge foncée. En outre, le couvert forestier n'est pas influencé par les changements saisonniers. Ainsi les valeurs radiométriques de ces zones ne varient pas dans les images satellites de la série temporelle. Tandis que

le graphe d'évolution de la figure 5.8 correspond à une entité spatio-temporelle de la classe *surface en eau* couverte par de la végétation aquatique. La surface en eau présente une couleur bleu foncé et la végétation apparait en rouge. Cette combinaison de valeur radiométrique est similaire à celle du couvert forestier. Ainsi les trois graphes d'évolution sont caractérisés par une dynamique similaire d'où leur regroupement dans le même cluster.

La Figure 5.11 et 5.12 représentent deux graphes d'évolution de la Basse Plaine de l'Aude et le Pic Saint Loup respectivement. Les deux graphes d'évolution sont regroupés dans le cluster 6. Ils illustrent l'évolution d'entités spatio-temporelles de la classe *végétation arbustive*. L'évolution de la végétation arbustive est progressive, voire constante au cours de temps. Ainsi la couleur des objets correspondant aux deux graphes d'évolution ne varie pas dans les images satellites de la série. Les objets de ces deux graphes sont caractérisés par les deux couleurs, rouge (végétation) et bleu (sol nu) correspondant un couvert végétal peu dense.

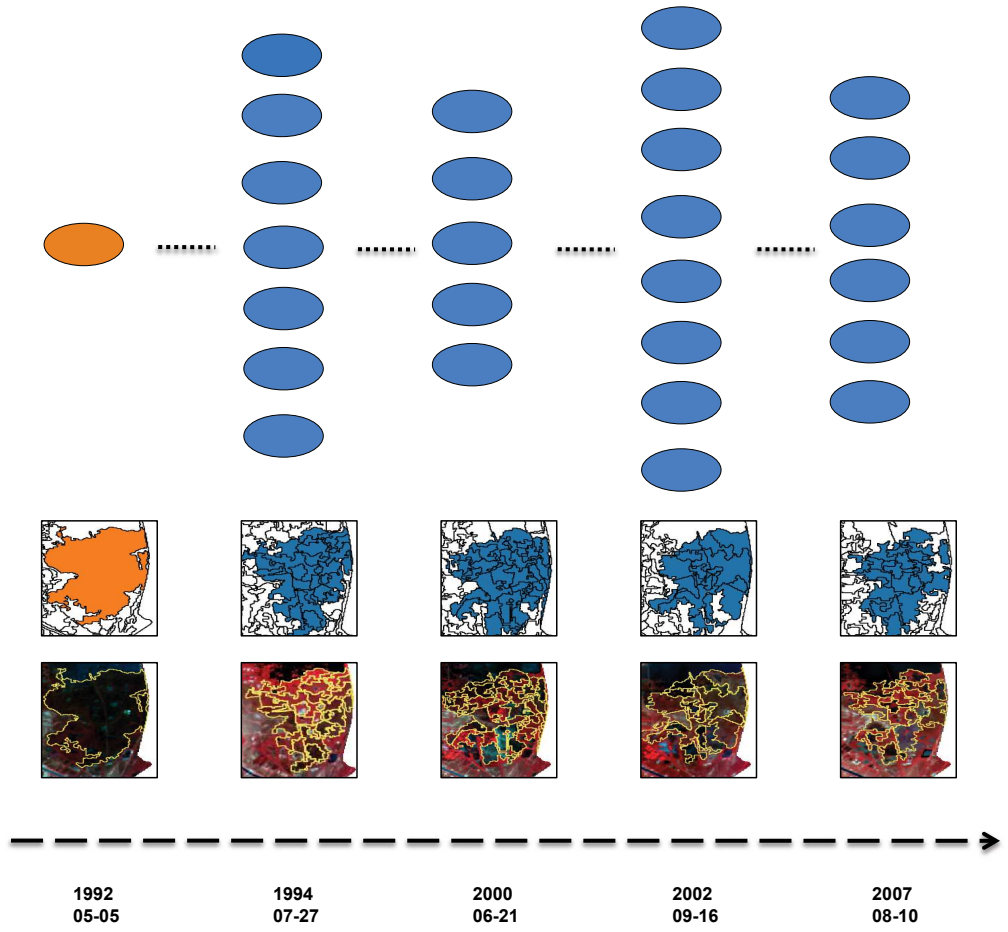


FIGURE 5.8 – Exemple d'un graphe d'évolution identifié sur la Basse Plaine de l'Aude qui représente l'évolution d'une *surface en eau* couverte par de la végétation aquatique regroupé dans le cluster 4.

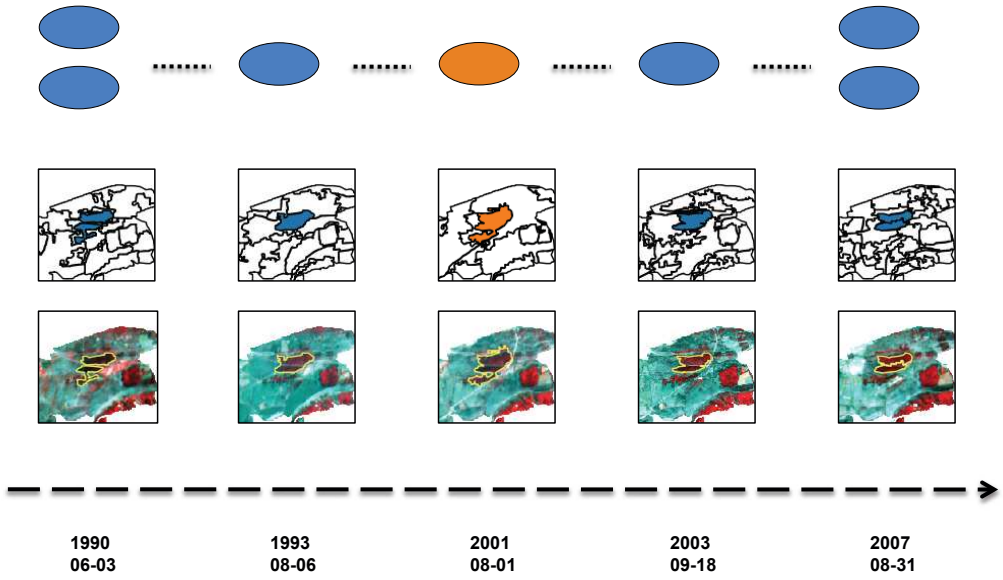


FIGURE 5.9 – Exemple d'un graphe d'évolution identifié sur la MMCA qui représente l'évolution d'une surface de *forêt* regroupé dans le cluster 4.

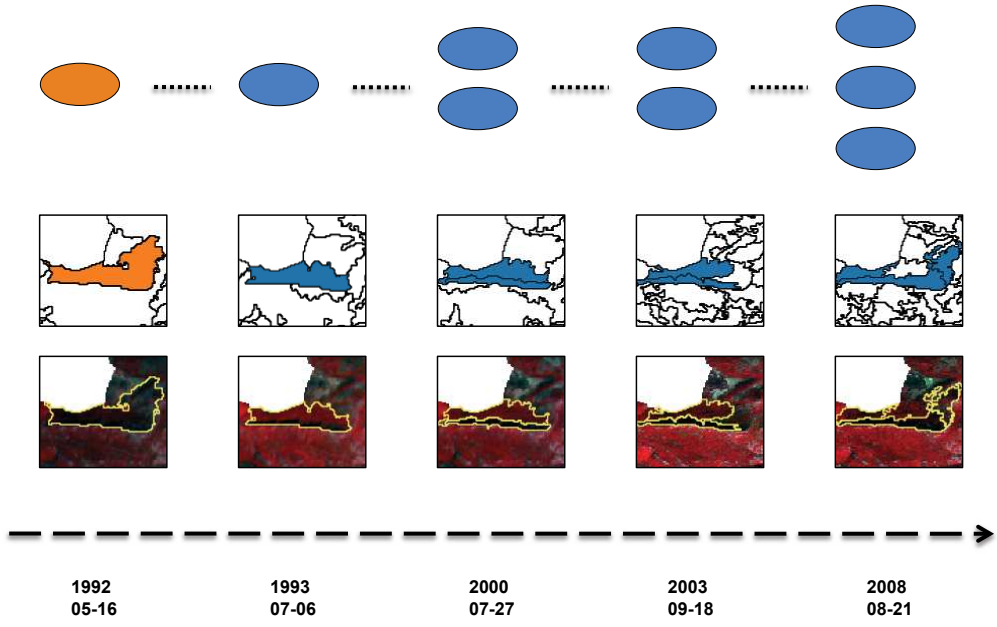


FIGURE 5.10 – Exemple d'un graphe d'évolution identifié sur le Pic Saint Loup qui représente l'évolution d'une surface de *forêt* regroupé dans le cluster 4.

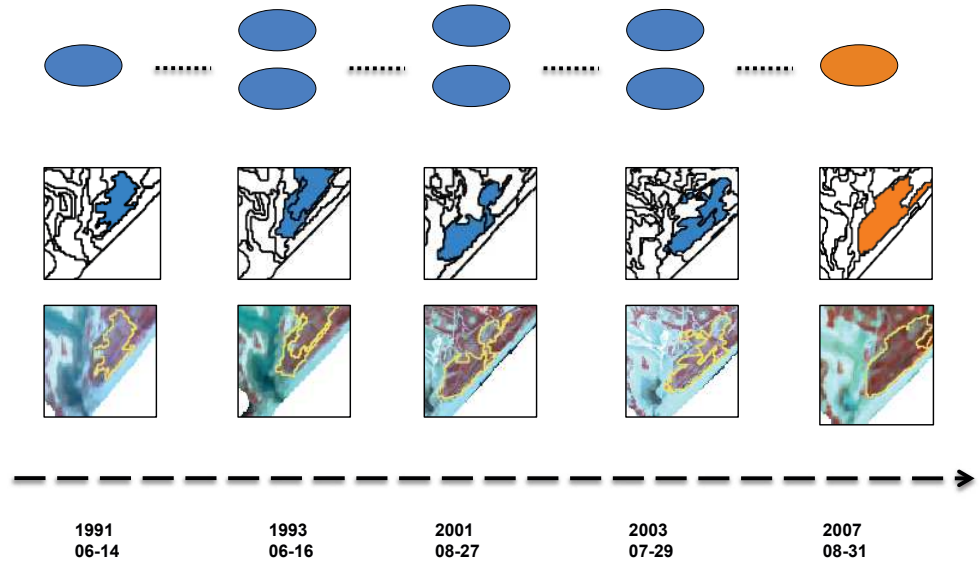


FIGURE 5.11 – Exemple d'un graphe d'évolution identifié sur la Basse Plaine de l'Aude qui représente l'évolution d'une surface couverte par de la *végétation arbustive* regroupé dans le cluster 6.

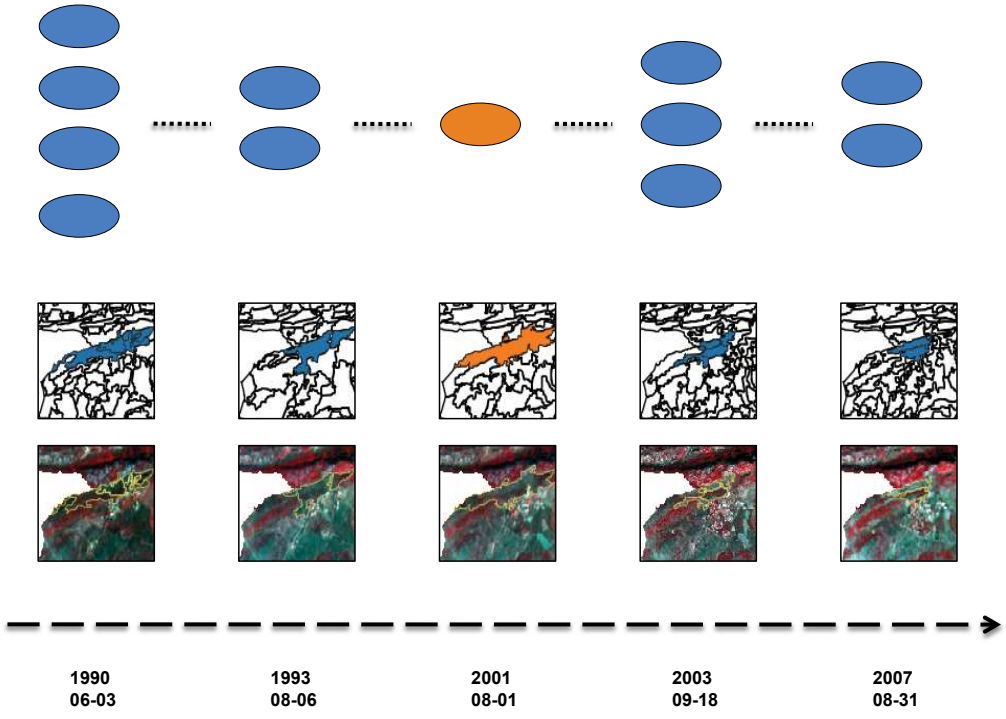


FIGURE 5.12 – Exemple d'un graphe d'évolution identifié sur le Pic Saint Loup qui représente l'évolution d'une surface couverte par de la *végétation arbustive* regroupé dans le cluster 6.

5.4 Conclusion

Ce chapitre présente une méthode d'analyse inter-sites exploitant des séries temporelles pluriannuelles. Notre objectif est l'identification des similarités inter-sites c'est-à-dire, des entités spatio-temporelles ayant une dynamique similaire appartenant à des sites différents. La méthode présentée dans le chapitre 4 a été adaptée aux spécificités des données. La méthode proposée permet d'identifier des entités spatio-temporelles dans des séries temporelles d'images satellites pluriannuelles. Les séries temporelles analysées sont segmentées puis leurs objets sont analysés afin d'identifier des objets de référence. L'évolution des entités spatio-temporelles est illustrée en utilisant des graphes d'évolution. Les graphes d'évolution sont comparés en utilisant des mesures de distance adéquates telle que Dynamic Time Warpping. Un algorithme de clustering est ensuite utilisé afin d'identifier des groupe de graphes d'évolution homogènes. En outre la méthode proposée permet d'accomplir une analyse inter-sites en combinant des entités spatio-temporelles identifiées sur plusieurs sites d'étude afin de mettre en évidence des similarités inter-sites.

Cette méthode a été évaluée sur trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en exploitant des séries temporelles acquissent sur une période de dix-huit ans de 1990 à 2008. Les résultats obtenus démontrent que la méthode proposée permet de mettre en lien les dynamiques de plusieurs sites d'étude en identifiant des entités spatio-temporelles dont les évolution sont similaires. En outre les entités spatio-temporelles ont été caractérisées par l'information radiométrique dont les bandes spectrales et indices radiométriques. La combinaison des bandes spectrales et des indices radiométriques permet une meilleure discrimination des entités spatio-temporelles et de leur évolution.

Bien que les résultats obtenus soit cohérents, le partitionnement des entités spatio-temporelles identifiées a été comparé à une classification experte. Ce partitionnement correspond à la classification experte. En effet, la classification experte est subjective et dépend de l'objectif de l'analyse des images satellites. Afin d'adapter le partitionnement automatique aux attentes de l'expert, nous allons proposer d'introduire des connaissances expertes dans le processus de clustering en utilisant des méthodes de clustering semi-supervisées. Le chapitre 6 concernera ainsi l'analyse semi-supervisée des séries temporelles d'images satellites.

Chapitre 6

Clustering semi-supervisé des séries temporelles

Sommaire

6.1 Introduction

Dans les Chapitres 4 et 5, nous avons présenté une méthode non supervisée pour l'analyse de séries temporelles d'images satellites annuelles et pluriannuelles respectivement. Cette méthode exploite des séries temporelles d'images satellites segmentées afin d'identifier des entités spatio-temporelles d'intérêt. L'évolution de ces entités spatio-temporelles est représentée au cours du temps par un graphe, nommé graphe d'évolution. Ces graphes d'évolution sont ensuite transformés en synospis. Les synopsis constituent une représentation synthétique des graphes d'évolution. Cette représentation permet d'appliquer un algorithme de clustering afin de regrouper les entités spatio-temporelles évoluant similairement. Cette méthode a été validée sur nos différents sites d'étude. Les résultats obtenus ont démontré que la représentation par synopsis permet d'identifier et de mettre en évidence des types d'évolution caractéristiques propres à chaque site d'étude.

Cependant, il existe pour un ensemble d'entités spatio-temporelles plusieurs partitionnements possible. Chacun des partitionnements est adapté à une des dimensions des données et à une tâche particulière. Afin de réaliser le partitionnement souhaité par l'utilisateur, nous proposons d'introduire des données étiquetés dans le processus de clustering pour le guider. Nous réalisons ainsi une analyse semi-supervisée des données.

Le choix des entités à sélectionner est une étape critique. Dans ce contexte, nous proposons une méthode de sélection nommée *CSEC* (Constraint Selection using an Ensemble of Clustering) basée sur un ensemble de partitionnements. Cette méthode identifie et sélectionne les entités les plus susceptibles d'être mal regroupées. Étiqueter ces entités permet de guider et d'améliorer les performances du clustering.

Dans ce chapitre, nous introduisons le clustering semi-supervisé des séries temporelles d'images satellites. Nous présentons le schéma général de la méthode (Section 6.2), cette méthode compte trois étapes. La première étape consiste à générer plusieurs partitionnements, la deuxième permet de créer une matrice de co-occurrence et enfin la dernière étape permet de sélectionner les contraintes. Nous présentons aussi le protocole d'évaluation (Section 6.3), puis nous rapportons et discutons les résultats obtenus sur nos sites d'étude (Section 6.4) .

6.2 Analyse semi-supervisée des STIS

Dans cette section, nous présentons notre méthode semi-supervisée pour l'analyse de séries temporelles d'images satellites et nous décrivons chacune des ses étapes en détail. Nous définissons les différents types de contraintes ainsi que l'algorithme de clustering par contraintes utilisé.

6.2.1 Vue globale de la méthode

L'analyse semi-supervisée de séries temporelles d'images satellites compte la phase de détection des entités spatio-temporelles ainsi que l'illustration de leurs évolutions par des graphes d'évolution puis des synopsis (Figure 6.1(Étape 1)). Ces deux étapes ont été définies dans les Chapitres 4 et 5. Nous proposons dans ce chapitre d'introduire les connaissances expertes dans le processus de clustering sous forme de contraintes (clustering par contraintes).

Les connaissances expertes visent à guider le processus de clustering des entités spatio-temporelles. Ces connaissances expertes se présentent sous forme de contraintes **Must-link** et **Cannot-link**. Les contraintes **Must-link** spécifient que deux entités spatio-temporelles possèdent une évolution similaire tandis que les contraintes **Cannot-link** spécifient que deux entités spatio-temporelles possèdent des évolutions différentes.

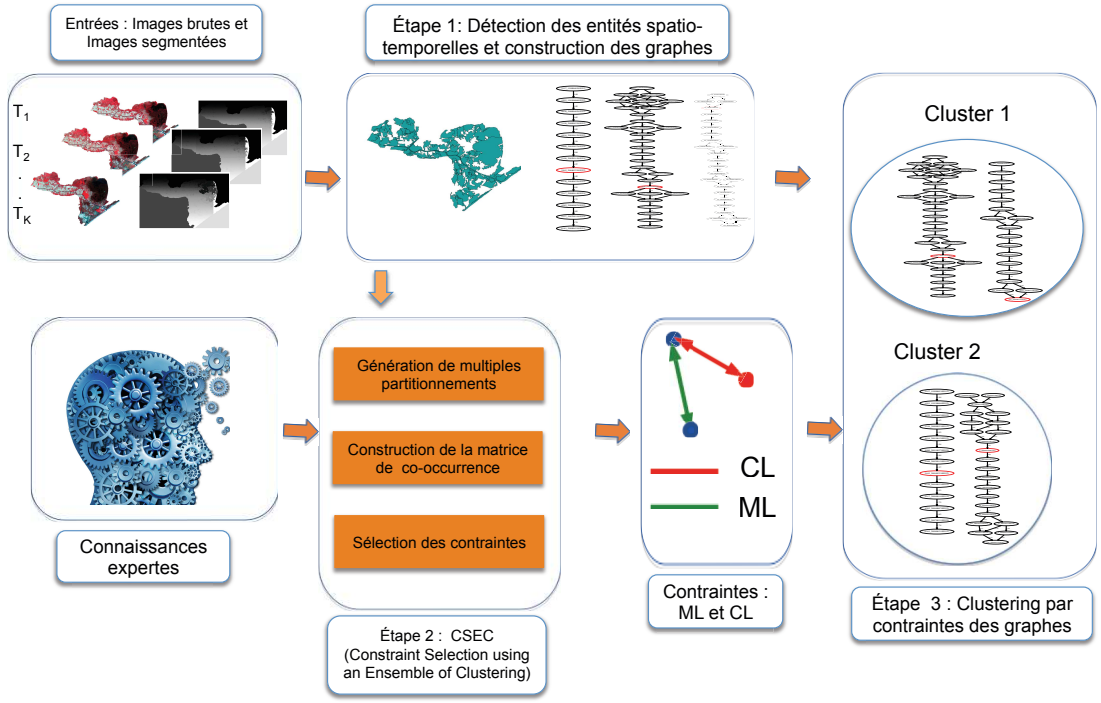


FIGURE 6.1 – Schéma général illustrant les différentes étapes de l'analyse semi-supervisée des entités spatio-temporelles.

Le clustering par contraintes (Figure 6.1(Étape 3)) implique la sélection de contraintes. Nous proposons une méthode de sélection de contraintes (Figure 6.1(Étape 2)) qui permet d'identifier les paires d'entités spatio-temporelles les plus informatives à fournir à l'utilisateur. Notre hypothèse se base sur la sélection des entités spatio-temporelles les plus susceptibles d'être mal regroupées et qui diminue les performance du clustering. En étiquetant ces instances, on guide l'algorithme

de clustering pour les regrouper correctement. Afin d'identifier ces entités spatio-temporelles, nous introduisons une méthode en trois étapes :

— (i) Génération de multiples partitionnements : Les données sont analysées par différents algorithmes de clustering, afin de générer plusieurs partitionnements.

— (ii) Construction de la matrice de co-occurrence : Les différents partitionnements générés permettent de calculer un score de co-occurrence pour chaque paire d'entités spatio-temporelles. Ce score indique quelles sont les entités les plus susceptibles d'être mal classées.

— (iii) Sélection des contraintes : En se basant sur les scores de co-occurrence, on sélectionne des paires d'entités spatio-temporelles considérées profitables pour l'algorithme de clustering. Ces paires d'entités sont fournies à l'expert afin de les étiqueter comme des contraintes **Must-link** ou **Cannot-link**.

6.2.2 Contraintes

Les contraintes permettent d'exprimer la connaissance experte. Ces contraintes sont exploitées par des algorithmes d'apprentissage semi-supervisé. Dans la littérature on définit principalement quatre types de contraintes divisés en deux groupes : les contraintes sur les entités (Basu et collab., 2008) et les contraintes sur les clusters (Davidson et Basu, 2007) :

— Les contraintes de type **Must-link** (ML) : Les contraintes **Must-link** sont spécifiées sur des paires d'entités E_i et $E_{i'}$. Elles indiquent que les deux entités sont similaires et doivent appartenir au même cluster, $L_{E_i} = L_{E_{i'}}$.

— Les contraintes **Cannot-link** (CL) : Les contraintes **Cannot-link** sont aussi spécifiées sur des paires d'entités $(E_i, E_{i'})$. Elles indiquent que les deux entités sont différentes et doivent être regroupées dans des clusters différents, $L_{E_i} \neq L_{E_{i'}}$.

— δ**-Constraint** : Les contraintes δ sont spécifiées sur des clusters. Elles indiquent la séparabilité minimale entre les entités de chaque paire de clusters. Ainsi, la distance entre chaque paire d'entités appartenant à deux clusters différents doit être supérieure à δ. Soit un partitionnement $P = \{C_1, C_2, ..., C_k\}$ et un ensemble d'entités $\mathbb{E} = \{E_1, E_2, ..., E_n\}$, la contrainte δ**-Constraint** est notée :

$$\forall\ E_i \in C_j, \forall\ E_{i'} \in C_{j'}\ \ alors\ \ Dist(E_i, E_{i'}) \geq \delta$$

— ϵ**-Constraint** : Les contraintes ϵ sont spécifiées sur des clusters. Elles indiquent la compacité minimale de chaque cluster. Cette contrainte définit une distance minimale ϵ entre chaque paire d'entités appartennat à un même cluster. Soit un partitionnement $P = \{C_1, C_2, ..., C_k\}$ et un ensemble d'entités $\mathbb{E} = \{E_1, E_2, ..., E_n\}$, la contrainte ϵ**-Constraint** est notée :

$$\forall \ E_i \in C_j, \ \ \exists \ E_{i'} \in C_j \ \ tel \ que \ \ Dist(E_i, E_{i'}) \leq \epsilon$$

Les contraintes **Must-link** et **Cannot-link** ont été introduites dans les travaux de (Wagstaff et Cardie, 2000). Elles représentent une forme de connaissance intuitive adaptable aux différents algorithmes de clustering existants (Wagstaff et Cardie, 2000). Simple à définir, les contraintes **Must-link** et **Cannot-link** permettent de spécifier les propriétés souhaitées pour le partitionnement. La méthode d'analyse de séries temporelles d'images satellites proposée se base sur ces deux types de contraintes.

La Figure 6.2 illustre les contraintes **Must-link** et **Cannot-link** ainsi que leur utilisation pour guider le processus de clustering. Les entités illustrées admettent deux regroupements différents en considérant l'attribut 1 ou l'attribut 2. Les contraintes **Must-link** et **Cannot-link** spécifient que le regroupement souhaité par l'utilisateur est celui considérant l'attribut 1.

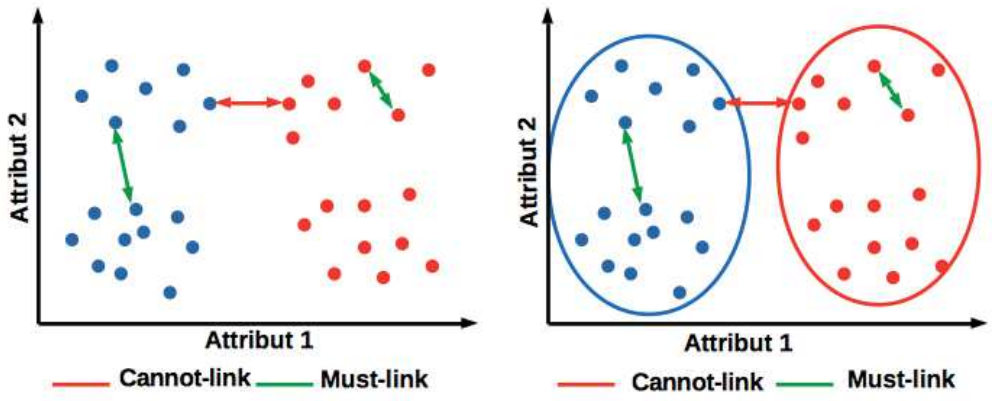


Figure 6.2 – Illustration des contraintes **Must-link** et **Cannot-link**(Basu et collab., 2008).

Les contraintes **Must-Link** et **Cannot-Link** partagent des propriétés intéressantes (Davidson et Basu, 2007). Les contraintes **Must-Link** définissent une relation symétrique et transitive :

$$\forall \ E_i, E_{i'}, E_{i''} \in \mathbb{E} \ \ tels \ que \ \ ML(E_i, E_{i'}) \ \ et \ \ ML(E_{i'}, E_{i''}) \Rightarrow ML(E_i, E_{i''})$$

Les contraintes **Cannot-link** définissent une relation symétrique mais non transitive. Cependant, il existe une relation de transitivité entre les contraintes **Must-link** et **Cannot-link** :

$$\begin{aligned} &\forall \ E_i, E_{i'}, E_{i''} \in \mathbb{E}, \ \ tels \ que \ \ (ML(E_i, E_{i'}) \ \ et \ \ CL(E_{i'}, E_{i''})) \\ &ou \ \ (CL(E_i, E_{i'}) \ \ et \ \ ML(E_{i'}, E_{i''})) \Rightarrow CL(E_i, E_{i''}) \end{aligned}$$

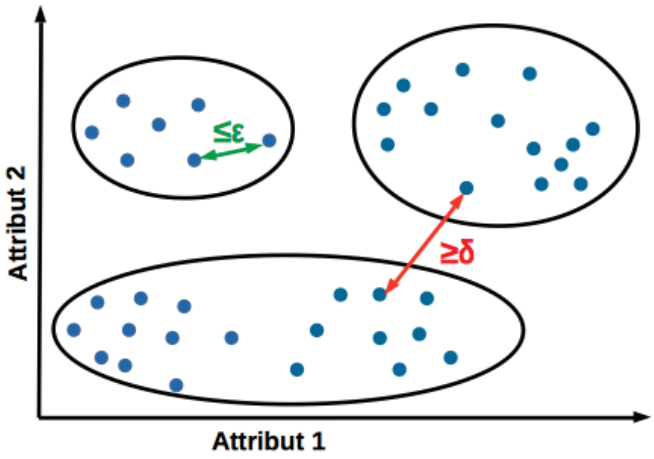


FIGURE 6.3 – Illustraction des contraintes **Epsilon** (ϵ) et **Delta** (δ) (DAVIDSON et BASU, 2007).

La Figure 6.3 illustre les contraintes δ et ϵ. Les contraintes δ permettent d'obtenir deux groupes d'entités avec une distante minimale égale à δ ce qui est équivalent à une conjonction de contraintes **Must-link** entre toutes les paires d'entités dont la distance est inférieure à δ et une conjonction de contraintes **Cannot-link** entre toutes les paires d'entités dont la distance est supérieure à δ. Les contraintes ϵ spécifient que pour chaque entité dans un cluster, il existe un voisinage dont la distance est inférieure à ϵ ce qui correspond à une disjonction de contraintes **Must-link** (DAVIDSON et RAVI, 2005a).

6.2.3 Génération de multiples partitionnements

La génération de multiples partitionnements est inspirée des méthodes de clustering par ensemble. Le clustering par ensemble consiste à générer un partitionnement final des données en combinant plusieurs partitionnements différents de ces mêmes données. Il a pour objectif d'améliorer le partitionnement individuel des données (HANSEN et SALAMON, 1990). Il compte deux étapes : l'étape de génération des différents partitionnements et l'étape de consensus qui permet de générer le partitionnement final (VEGA-PONS et RUIZ-SHULCLOPER, 2011). La Figure 6.4 illustre le processus de clustering par ensemble.

Les partitionnements peuvent être générés de plusieurs manières (VEGA-PONS et RUIZ-SHULCLOPER, 2011), parmi celle-ci :

— Utiliser différents algorithmes de clustering.

— Utiliser le même algorithme de clustering en variant ses paramètres.

— Projeter les données dans différents espaces de représentation.

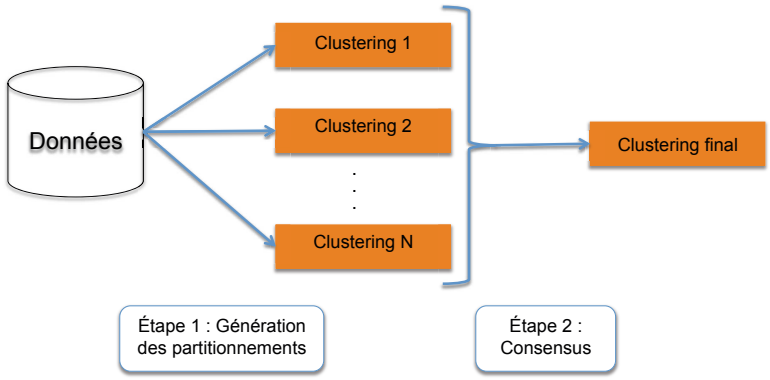


FIGURE 6.4 – Processus général du clustering par ensemble.

— Sélectionner différents sous-ensembles d'attributs des données.

Nous avons réalisé un ensemble de partitionnements sur les données en utilisant différents algorithmes de clustering. La première étape est la sélection des N algorithmes de clustering à utiliser. Une fois sélectionnés, chaque algorithme est appliqué sur les données afin de générer un partitionnement. On obtient N partitionnements différents de données, $\mathbb{P} = \{P_1, P_2, ..., P_N\}$. Chaque partitionnement P_j correspond a un ensemble de cluster $P_j = \{C_1, C_2, ..., C_k\}$. Pour chaque instance de données E_i, on identifie un label (cluster) pour chaque partitionnement P_j noté, $L_{E_i}^{P_j}$.

6.2.4 Construction de la matrice de co-occurrence

Afin d'identifier les instances ambiguës qui doivent être étiquetées, on calcule un score de confiance de regroupement. On considère que deux entités similaires sont regroupées toujours ensemble dans les différents partitionnements tandis que les entités ambiguës sont regroupées de manière confuse par les différents partitionnements, c'est-à-dire certains partitionnements les regroupent dans le même cluster tandis que d'autres les regroupent différemment.

D'abord, on identifie pour chaque paire d'entités $(E_i, E_{i'})$, les partitionnements qui les regroupent dans le même cluster.

$$P(E_i, E_{i'}) = \{P_j,\ L_{E_i}^{P_j} = L_{E_{i'}}^{P_j}\} \tag{6.1}$$

On analyse les différentes paires d'entités. Nous considérons que deux entités regroupées dans le même cluster dans plus de deux partitionnements $P(E_i, E_{i'}) \geq 2$, sont similaires et non ambiguës. Plus les deux instances sont regroupées ensemble par les différents partitionnements plus le score de confiance des deux entités doit être élevé. Ainsi, on identifie les paires de partitionnements, $(P_j, P_{j'})$, qui regroupent les deux entités $(E_i, E_{i'})$ dans un même cluster.

$$P'(E_i, E_{i'}) = \{(P_j, P_{j'})\ \ tel\ que\ \ L_{E_i}^{P_j} = L_{E_{i'}}^{P_j}\ \ et\ \ L_{E_i}^{P_{j'}} = L_{E_{i'}}^{P_{j'}}\} \tag{6.2}$$

On définit, pour chaque paire d'entités $(E_i, E_{i'})$, un score de confiance de partitionnement, nommé score de co-occurrence $W(E_i, E_{i'})$, égal au nombre de paires de partitionnement qui les regroupent ensemble :

$$W(E_i, E_{i'}) = |\{(P_j, P_{j'}) \ \ telle\ que \ \ L_{E_i}^{P_j} = L_{E_{i'}}^{P_j} \ \ et \ \ L_{E_i}^{P_{j'}} = L_{E_{i'}}^{P_{j'}}\}| \tag{6.3}$$

Afin de pondérer ce score, on compare chaque paire de partitionnement $(P_i, P_{i'})$, et on calcule leur similarité globale en utilisant l'indice d'information mutuelle normalisée (Section 2.2.4.2). Cet indice permet de quantifier le degré d'agrément entre les deux partitionnements.

$$Sim(P_j, P_{j'}) = NMI(P_j, P_{j'}) \tag{6.4}$$

En sommant la similarité entre les paires de partitionnements, on obtient un score de co-occurrence comme défini par l'équation 6.5 :

$$\begin{aligned} W'(E_i, E_{i'}) =& |\{(P_j, P_j) \ \ telle\ que \ \ L_{E_i}^{P_j} = L_{E_{i'}}^{P_j} \ \ et \ \ L_{E_i}^{P_{j'}} = L_{E_{i'}}^{P_{j'}}\}| \\ &+ \sum_{P_j, P_{j'} \in P'(E_i, E_{i'})} NMI(P_j, P_{j'}) \end{aligned} \tag{6.5}$$

Dans cette étape, nous définissons une matrice de co-occurrence comprenant les scores de co-occurrences calculés entre chaque paires d'entités.

6.2.5 Sélection des contraintes

Les contraintes permettent de formaliser la connaissance de l'expert qui est intégrée dans le processus de clustering. Notre hypothèse est de sélectionner les entités les plus susceptibles d'être mal classées. On se basant sur les scores de co-occurrence précédemment calculés, on identifie les instances à sélectionner pour interroger l'expert. Les paires d'instances sont ordonnées selon leur score de co-occurrence pour sélectionner celles dont le score est minimal.

$$Const = \{(E_i, E_{i'}) \in \mathbb{E} \ \ telle\ que \ \ W'(E_i, E_{i'}) = argmin_{(E_i, E_{i'}) \in \mathbb{E}} W'(E_i, E_{i'})\} \tag{6.6}$$

6.2.6 Clustering des données

Les contraintes constituent les connaissances expertes qui permettrons de réaliser un partitionnement de données adapté aux besoins de l'utilisateur. Les différents algorithmes de clustering de la littérature ont été modifiés pour intégrer les contraintes dans leur processus de partitionnement. Parmi ces algorithmes, nous pouvons citer : le Kmeans (Bilenko et collab., 2004 ; Wagstaff et collab., 2001), le DBSCAN

(Lelis et Sander, 2009 ; Ruiz et collab., 2007), le clustering hiérarchique (Davidson et Ravi, 2005b), le clustering spectral (Wang et Davidson, 2010), le fuzzy c-means (Grira et collab., 2006) et le clustering leader (Vu et Bouchon-Meunier, 2009). Ces différents algorithmes utilisent les contraintes de différentes manières (Davidson et Basu, 2007) :

— Modification de la phase d'initialisation des clusters ;

— Modification de la fonction objectif ;

— Modification de l'affectation des entités aux cluster ;

— Modification de l'espace métrique des données.

Dans la phase de clustering nous nous somme basé sur l'algorithme Kmeans. L'Algorithme 4 décrit le partitionnement Kmeans (MacQueen et collab., 1967). Le Kmeans est l'un des algorithmes les plus connus et utilisés en littérature (Xu et Wunsch, 2005). Cet algorithme permet de partitionner un ensemble d'entités, $\mathbb{E} = \{E_1, E_2, ..., E_n\}$, en k clusters. On note $P = \{C_1, C_2, ..., C_k\}$ le partitionnement résultat. Ce partitionnement est réalisé d'une manière itérative, le regroupement des entités est modifié à chaque itération dans l'objectif de minimiser la distance intra-cluster et maximiser la distance inter-cluster.

Cet algorithme compte 3 étapes :

— L'initialisation (Ligne 4) permet de déterminer les centres des clusters, $\mu = \{\mu_1, \mu_2, ..., \mu_k\}$, en choisissant k entités aléatoirement ;

— Le partitionnement permet d'affecter chaque entité à un cluster (Ligne 6-8). Afin d'identifier ce cluster, nous calculons la distance entre l'entité et chacun des centres des clusters puis nous sélectionnons celui dont la distance est minimale ;

— La mise à jour des centres (Ligne 11-12) permet de réajuster les centres, en calculant pour chaque cluster la moyenne de ses entités.

Les étapes 2 et 3 sont répétées jusqu'à convergence de la fonction objective suivante (Arthur et Vassilvitskii, 2007) :

$$J_{Kmeans} = \frac{1}{2} \sum_{j=1}^{k} \sum_{E_i \in C_j} \left\| (E_i - \mu_j) \right\|^2$$

Algorithme 4 : Kmeans

Entrées : Données $\mathbb{E}= \{E_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre de clusters k
Sorties : *Partionnement* $P = \{C_1, C_2, .., C_k\}$
(1).Initialisation des centres des clusters $\mu = \{\mu_1, \mu_2, ..., \mu_k\}$
(2).Affectation des entités ;
pour tout $E_i \in \mathbb{E}$ **faire**
    Affectation de l'entité E_i au cluster $C_j \in P$ tel que :
        $Dist(E_i, \mu_j) = argmin_{j=1,k} \left\|E_i - \mu_j\right\|^2$
fin pour
(3).Mise à jour des centres
pour tout $\mu_i \in \{\mu_j\}_{j=1}^{k}$ **faire**
    $\mu_j = \frac{1}{|C_j|} \sum_{E_i \in C_j} E_i$
fin pour
Répéter (2) et (3) jusqu'à convergence
retourner $P = \{C_1, C_2, .., C_k\}$

Afin de pouvoir intégrer les contraintes dans le processus de clustering, nous avons utilisé une variante du Kmeans. De nombreux travaux de la littérature ont proposé des variantes de cette algorithme pour prendre en compte des contraintes. Nous citons l'algorithme COP-Kmeans (Wagstaff et collab., 2001), l'agorithme PCKmeans (Basu et collab., 2004) et l'agorithme MPCKmeans (Metric Pairwise Constraints Kmeans) (Bilenko et collab., 2004).

Nous avons opté pour le MPCKmeans. Cet algorithme intègre les contraintes dans le processus de partitionnement. Les contraintes sont utilisées pour l'apprentissage de métriques. Elles permettent d'entrainer une mesure de distance propre pour chaque cluster. Son objectif est d'apprendre une nouvelle distance qui permet de minimiser la distance entre les paires d'entités **Must-link** et maximiser la distance entre les paire d'entités **Cannot-link**. Le MPCKmeans introduit une nouvelle fonction objective à optimiser, définie par les Équations 6.7, 6.8 et 6.9 :

$$J_{MPCKmeans} = \sum_{j=1}^{k} \sum_{E_i \in C_j} (\left\|E_i - \mu_j\right\|_{A_j}^2 - log(det(A_j))) \tag{6.7}$$

$$+ \sum_{E_i, E_{i'} \in ML} w_{i,i'} f_{ML}(E_i, E_{i'}) \mathbb{1}[L_{E_i} \neq L_{E_{i'}}] \tag{6.8}$$

$$+ \sum_{E_i, E_{i'} \in CL} \bar{w}_{i,i'} f_{CL}(E_i, E_{i'}) \mathbb{1}[L_{E_i} = L_{E_{i'}}] \tag{6.9}$$

Cette fonction objective est définie par la somme des distances intra-cluster (Équation 6.7), le coût de violation des contraintes **Must-link** (Équation 6.8) et le coût de violation des contraintes **Cannot-link** (Équation 6.9) (Van Craenendonck et Blockeel, 2017).

La distance intra-cluster est paramétrée par une matrice symétrique positive A_j définie pour chaque cluster C_j. Elle est définie par l'Équation 6.10 :

$$A_j = |C_j|(\sum_{E_i \in C_j} (E_i - \mu_j)(E_i - \mu_j)^T$$
$$+ \sum_{E_i, E_{i'} \in ML_j} \frac{1}{2} w_{i,i'} (E_i - E_{i'})(E_i - E_{i'})^T \mathbb{1}[L_{E_i} \neq L_{E_{i'}}]$$
$$+ \sum_{E_i, E_{i'} \in CL_j} \bar{w}_{i,i'} ((E'_j - E''_j)(E'_j - E''_j)^T - (E_i - E_{i'})(E_i - E_{i'})^T) \mathbb{1}[L_{E_i} = L_{E_{i'}}])^{-1} \tag{6.10}$$

Où :

— (E'_j, E''_j) représente la paire d'entités appartenant au cluster C_j dont la distance est maximale ;

— ML_j représente l'ensemble de contraintes **Must-link** incluses dans le cluster C_j ;

— CL_j représente l'ensemble de contraintes **Cannot-link** incluses dans le cluster C_j.

Le coût de violation des contraintes est défini par un poids $w_{i,i'}$ et $\bar{w}_{i,i'}$ ainsi qu'une fonction f_{ML} et f_{CL} de pondération pour les contraintes **Must-link** et **Cannot-link** respectivement.

L'algorithme 5 décrit le partionnment MPCKmeans (Bilenko et collab., 2004). Le MPCKmeans étant une variante de Kmeans, il compte les mêmes étapes que l'algorithme Kmeans :

— L'initialisation est réalisée en créant un ensemble de voisinages $\{Ne_p\}_{p=1}^{\lambda}$ ou groupes d'entités liées par des contraintes **Must-link**. Pour chaque groupe, nous calculons un centre en moyennant ses entités. Si le nombre de groupes est inférieur à k, le reste des centres sont initialisés aléatoirement (Ligne7-13) ;

— Le partitionnement est réalisé en identifiant pour chaque entité un cluster. On sélectionne le cluster qui minimise la distance entre l'entité et son centre (Ligne 15-20). Le calcul de distance est basé sur la distance locale définie par la matrice A_j ainsi qu'un coût de violation des contraintes engendré par cette affectation ;

— La mise à jour des centres (Ligne 22-24) permet de réajuster les centres, en calculant pour chaque cluster la moyenne de ses entités.

Le MPCKmeans compte une étape supplémentaire qui est l'apprentissage de métriques. Il est réalisé en mettant à jour la matrice A_j (Ligne 26-31).

Les étapes 2, 3 et 4 sont répétées jusqu'à convergence de la fonction objective.

Algorithme 5 : MPCKmeans

Entrées : Données $\mathbb{E}= \{E_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre de clusters k
Contraintes **Must-link** : $\{ML(E_i, E_{i'})\} \subseteq \mathbb{R}^d * \mathbb{R}^d$
Contraintes **Cannot-link** : $\{CL(E_i, E_{i'})\} \subseteq \mathbb{R}^d * \mathbb{R}^d$
Sorties : *Partionnement* $P = \{C_1, C_2, .., C_k\}$
(1).Initialisation des centres des clusters $\mu = \{\mu_1, \mu_2, ..., \mu_k\}$;
Créer λ ensembles de voisinage : $\{Ne_p\}_{p=1}^{\lambda}$ à partir des contraintes ML et CL
si $\lambda > k$ **alors**
    Initialiser $\mu_j \in \{\mu\}$ par la moyenne des instances de $\{Ne_p\}_{p=1}^{\lambda}$ en utilisant la stratégie de la traversée la plus éloignée, en commençant par le plus grand ensemble de voisinage
sinon si $\lambda < k$ **alors**
    Initialiser $\mu_j \in \{\mu\}$ par la moyenne des instances de $\{Ne_p\}_{p=1}^{\lambda}$
    Initialiser le reste des centres aléatoirement
fin si
(2).Affectation des entités ;
pour tout $E_i \in \mathbb{E}$ **faire**
    Affectation de l'entité E_i au cluster $C_j \in P$ tel que :
        $Dist(E_i, \mu_j) = argmin_{j=1,k}(\|E_i - \mu_j\|^2_{A_j} - log(det(A_j))$
        $+ \sum_{E_i, E_{i'} \in ML} w_{i,i'} f_{ML}(E_i, E_{i'}) \mathbb{1}[j \neq L_{E_{i'}}]$
        $+ \sum_{E_i, E_{i'} \in CL} \bar{w}_{i,i'} f_{CL}(E_i, E_{i'}) \mathbb{1}[j = L_{E_{i'}}])$
fin pour
(3).Mise à jour des centres
pour tout $\mu_i \in \{\mu_j\}_{j=1}^{k}$ **faire**
    $\mu_j = \frac{1}{|C_j|} \sum_{E_i \in C_j} E_i$
fin pour
(4).Mise à jour de la distance
pour tout $A_j \in \{A_j\}_{j=1}^{k}$ **faire**
    $A_j = |C_j|(\sum_{E_i \in C_j} (E_i - \mu_j)(E_i - \mu_j)^T$
        $+ \sum_{E_i, E_{i'} \in ML_j} \frac{1}{2} w_{i,i'} (E_i - E_{i'})(E_i - E_{i'}) \mathbb{1}[L_{E_i} \neq L_{E_{i'}}]$
        $+ \sum_{E_i, E_{i'} \in CL_j} \bar{w}_{i,i'} (E'_j - E''_j)(E'_j - E''_j)^T$
        $-(E_i - E_{i'})(E_i - E_{i'}) \mathbb{1}[L_{E_i} = L_{E_{i'}}])^{-1}$
fin pour
Répéter (2),(3) et(4) jusqu'à convergence
retourner $P = \{C_1, C_2, .., C_k\}$

6.3 Expérimentation

Dans cette section, nous décrivons notre protocole expérimental dans l'objectif d'évaluer l'approche d'analyse d'entités spatio-temporelles utilisant le clustering semi-supervisé (par contraintes). Le protocole expérimental adopté permet de comparer l'analyse semi-supervisée à l'analyse non supervisée. Il permet aussi d'évaluer la méthode de sélection de contraintes *CSEC* en la comparant à la méthode de sélection de contraintes aléatoire. Par la suite, nous présentons et discutons nos résultats.

L'évaluation de la méthode proposée est réalisée sur nos quatre sites d'étude : La Basse Plaine de l'Aude, la Vallée du Libron, la Montagne de la Moure et Casse d'Amelas et le Pic Saint Loup.

6.3.1 Protocole expérimental

Afin d'évaluer notre méthode, nous avons soumis les entités spatio-temporelles à un expert du terrain afin de les classifier selon leur évolution. L'expert du terrain a réalisé une analyse visuelle puis a regroupé les entités spatio-temporelles ayant des évolutions similaires. La classification de l'expert a été utilisée pour valider les résultats de la méthode proposée et évaluer ses performances. Les deux regroupements sont comparés puis leur similarité est calculée, plus les deux partitionnements sont proches, meilleur sont les performances de notre méthode. La similarité entre les deux partitionnements a été mesurée en utilisant les indices de qualité externes NMI (Normalized Mutual Information) et ARI (Ajusted Rand Index) (Section 2.2.4.2).

Dans un premier temps, nous avons évalué les performances de la méthode d'analyse des STIS par rapport à la classification de l'expert. Dans un deuxième temps, nous avons comparé les performances de notre méthode à celles de l'état de l'art.

Nous avons comparé les performances de la méthode proposée à deux méthodes de base : (i) L'analyse non supervisée et (ii) l'analyse semi-supervisée basée sur de clustering par contraintes sélectionnées aléatoirement. L'analyse non supervisée est réalisée en utilisant un algorithme de clustering sans contraintes. Le clustering par contraintes aléatoires se base sur l'algorithme MPCKmeans, à la différence de la méthode proposé les contraintes sont sélectionnées d'une manière aléatoire.

L'objective est d'une part, de comparer les méthodes semi-supervisées aux méthodes non supervisées, et d'une autre part, de comparer les approches de sélection des contraintes : l'approche de sélection proposée et l'approche de sélection aléatoire.

La méthode proposée repose sur l'exploitation de contraintes pour guider le processus de clustering. Afin de déterminer le nombre de contraintes permettant d'améliorer les performances de clustering, nous avons réalisé des expérimentations génériques. Nous avons fait varier le nombre de contraintes dans l'intervalle $[0, 200]$ avec un pas de cinq contraintes, puis reporté les performances du clustering en terme des indices NMI et ARI.

Notre méthode de sélection de contraintes repose sur plusieurs résultats de clustering, afin de les réaliser, nous avons sélectionné trois algorithmes de clustering les plus utilisées dans la littérature : l'algorithme hiérarchique, l'algorithme spectral et l'algorithme kmeans (Tan et collab., 2005).

L'évaluation de notre approche a été réalisée sur des données réelles, correspondant à nos quatre sites d'étude : la Basse Plaine de l'Aude, la Vallée du Libron, la MMCA et le Pic Saint Loup. Nous avons exploité les séries temporelles d'images satellites Landsat décrivant la Basse Plaine de l'Aude et la Vallée du Libron sur une période de huit mois. Nous avons aussi exploité les séries temporelles d'images satellites Spot décrivant la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup sur une période de dix-huit ans.

6.4 Évaluation expérimentale

Nous avons évalué les performances de notre méthode semi-supervisée d'analyse de séries temporelles d'images satellites en utilisant des STIS Landsat et Spot. Nous avons reporté les résultats obtenus dans les Figures 6.5 à 6.8 en ce qui concerne les séries temporelles Landsat. Tandis que les résultats obtenus en considérant les séries temporelles Spot sont reportés dans les Figures 6.9 à 6.12

6.4.1 STIS Landsat

Les images Landsat décrivent la Basse Plaine de l'Aude et la Vallée du Libron. Les Figures 6.5 et 6.6 reportent les résultats ARI et NMI de l'analyse des entités spatio-temporelles identifiées sur la Basse Plaine de l'Aude. Chacune des figures comprend trois courbes correspondant aux trois méthodes qui sont : l'analyse non supervisée, l'analyse semi-supervisée basée sur une sélection de contraintes aléatoire et l'approche d'analyse semi-supervisée proposée nommée *CSEC*. Les résultats montrent que la méthode semi-supervisée proposée améliore les performance du clustering avec un valeur ARI égale à 0.5 comparée à 0.3 pour l'analyse non supervisée.

Les résultats s'améliorent en augmentant le nombre de contraintes sélectionnées. À partir de 40 contraintes, les performances de la méthode *CSEC* sont plus élevées que les performances du clustering sans contraintes. Contrairement aux résultats de *CSEC*, les performances de clustering semi-supervisé avec contraintes aléatoires diminuent en augmentant le nombre de contraintes.

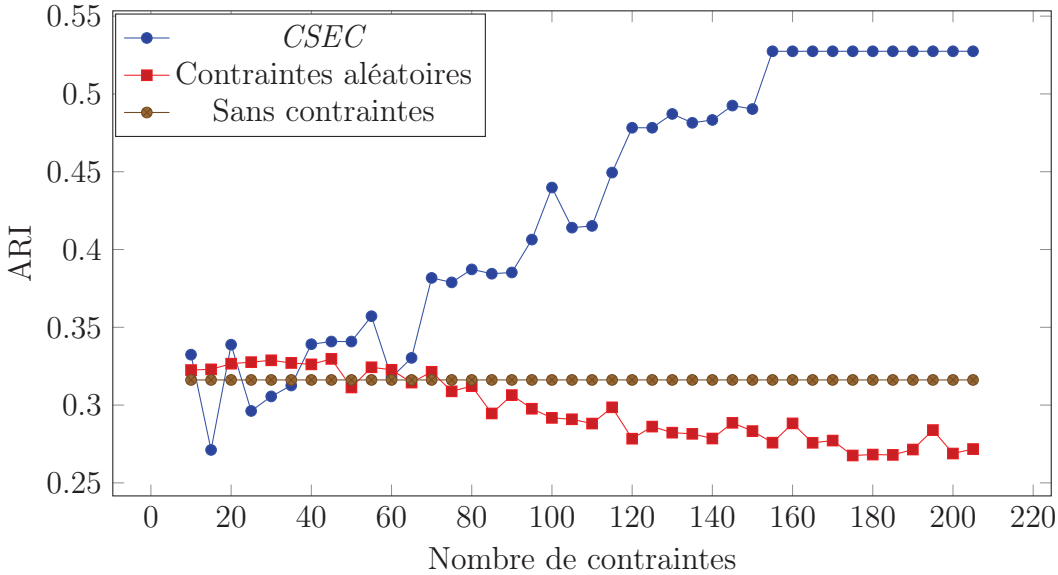


FIGURE 6.5 – Les résultats ARI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur la Basse Plaine de L'Aude.

De même que pour le ARI, les valeurs NMI obtenues montrent que la méthode

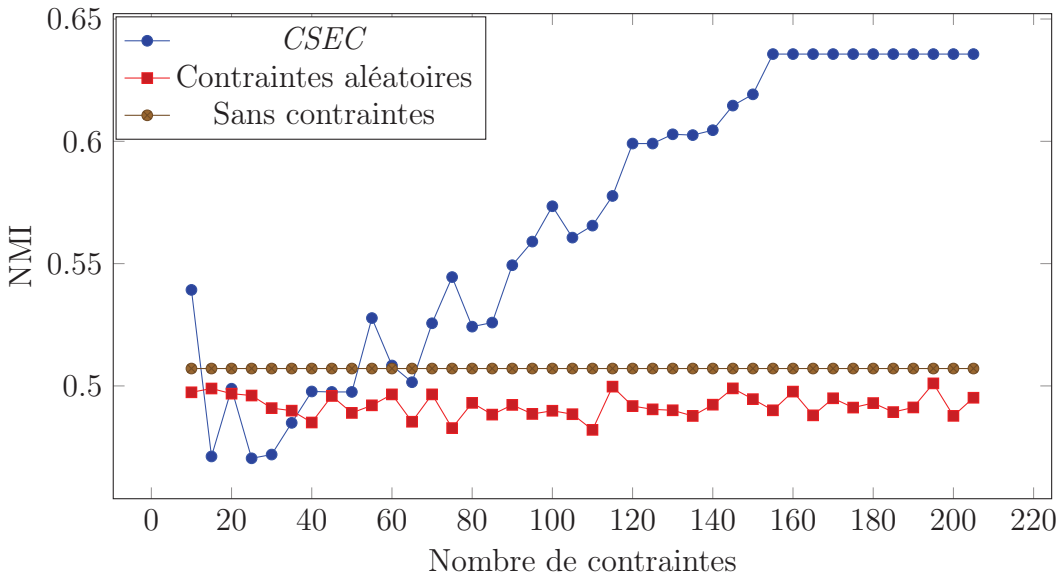


FIGURE 6.6 – Les résultats NMI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur la Basse Plaine de L'Aude.

CSEC est la plus performante des trois méthodes avec une valeur de 0.6 comparée à une valeur de 0.5 pour le clustering sans contraintes. Tandis que le clustering semi-supervisé avec contraintes aléatoires est la moins performante. Les valeurs NMI montrent que ses performances restent constantes malgré le nombre de contraintes qui augmente.

Les Figures 6.7 et 6.8 illustrent les résultats de l'analyse de la Vallée du Libron. Les valeurs ARI et NMI sur les deux figures montrent que les performances de la méthode *CSEC* sont les plus élevées. On note une valeur de 0.7 pour l'indice NMI et une valeur de 0.6 pour l'indice ARI comparées à des valeurs de 0.5 et 0.3 obtenues pour le clustering sans contraintes.

Considérant le clustering par contraintes aléatoires, contrairement aux résultats sur la Basse Plaine de L'Aude, les résultats obtenus pour la Vallée du Libron montrent qu'il est plus performant que les clustering sans contraintes.

6.4.2 STIS Spot

Les Figures 6.9 à 6.12 illustrent les résultats de l'analyse inter-site des trois zones : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup. Les entités spatio-temporelles sont d'abord caractérisées par les bandes spectrales (Figures6.9 et 6.10) puis par les bandes spectrales combinées aux indices radiométriques (Figure 6.11 et 6.12).

Les résultats ARI et NMI présentés dans les Figures 6.9 et 6.10 montrent que la méthode *CSEC* est la plus performante. La valeur ARI obtenue est égale à 0.3 et la valeur NMI est égale à 0.45. Tandis que pour le clustering sans contraintes,

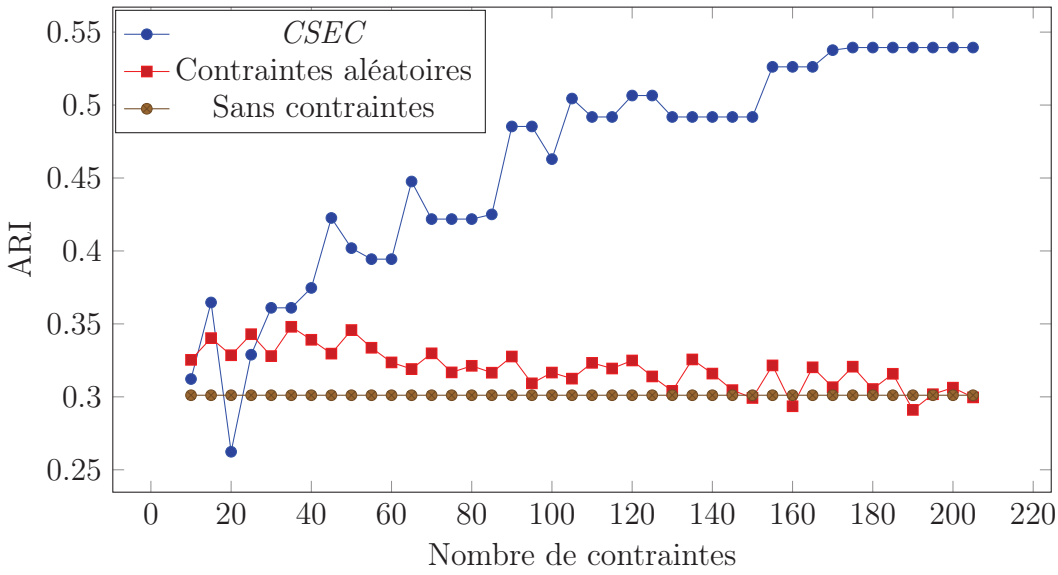


FIGURE 6.7 – Les résultat ARI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur la Vallée du Libron.

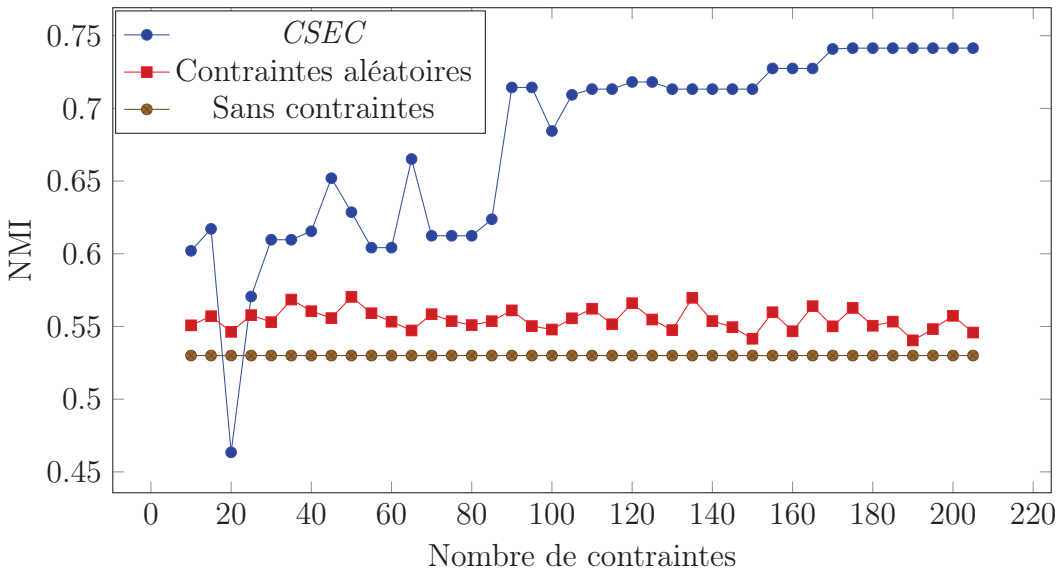


FIGURE 6.8 – Les résultat NMI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur la Vallée de Libron.

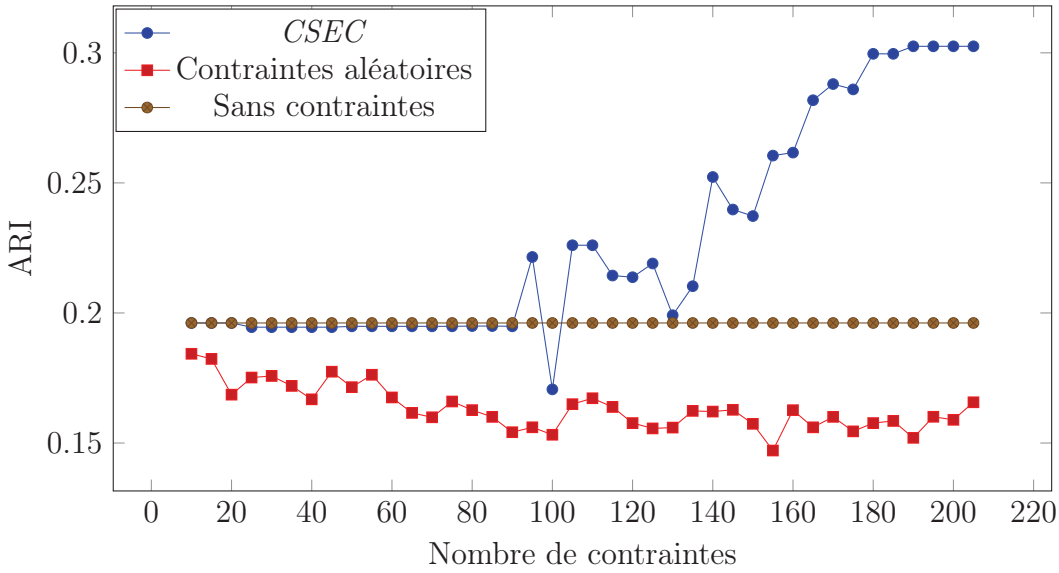


FIGURE 6.9 – Les résultats ARI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur les trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en utilisant les bandes spectrales.

nous obtentions une valeur de ARI égale à 0.2 et une valeur NMI égale à 0.36. Le clustering avec contraintes aléatoires obtient les performances les plus faibles.

Les résultats illustrés dans les Figures 6.11 et 6.12 montrent que la méthode *CSEC* est la plus performante. Elle obtient une valeur ARI égale à 0.4 comparée à 0.15 pour le clustering sans contraintes et une valeur NMI égale à 0.5 comparée à 0.35 pour le clustering sans contraintes. Le clustering avec les contraintes aléatoires quant à lui obtient les performance les plus faibles.

En considérant ces résultats, nous concluons que les bandes spectrales combinées aux indices radiométriques permettent une meilleurs discrimination entre les entités spatio-temporelles comparées aux bandes spectrales uniquement, ce qui appuie les résultats obtenus dans le Chapitre 5.

L'analyse réalisée sur les séries temporelles Landsat et Spot démontre que le clustering semi-supervisé par contraintes n'améliore pas les performances du clustering. En effet, en comparant les résultats du clustering par contraintes aléatoires au clustering sans contraintes, nous déduisons que les contraintes peuvent dégrader les performances du clustering (WAGSTAFF, 2006). Le choix des contraintes est une étape critique dans le processus du clustering supervisé, sélectionner des contraintes profitables au clustering permet d'améliorer le partitionnement comme le montre les résultats obtenus en utilisant la méthode de sélection des contraintes *CSEC*.

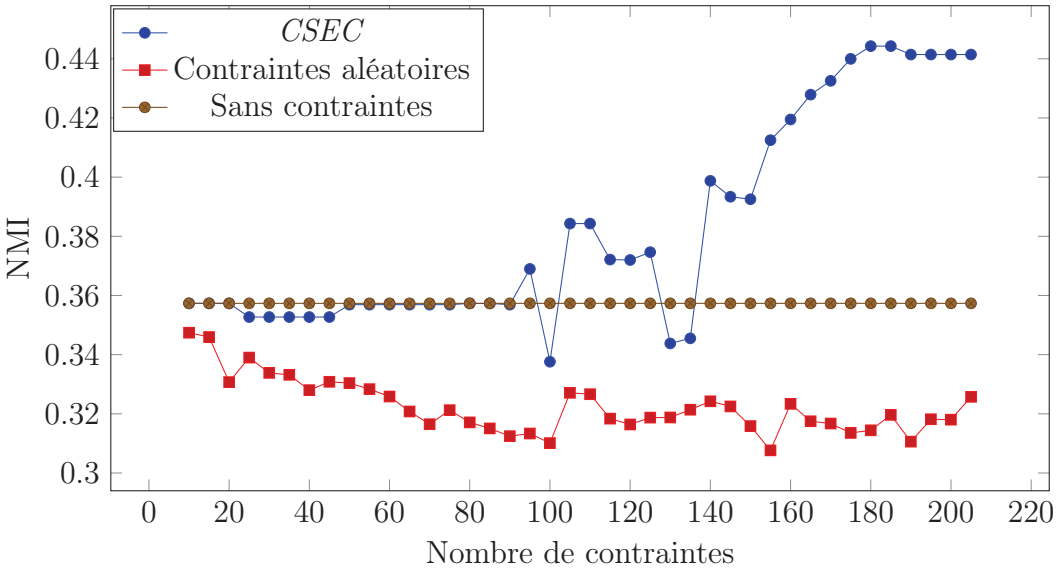


FIGURE 6.10 – Les résultats NMI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur les trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en utilisant les bandes spectrales.

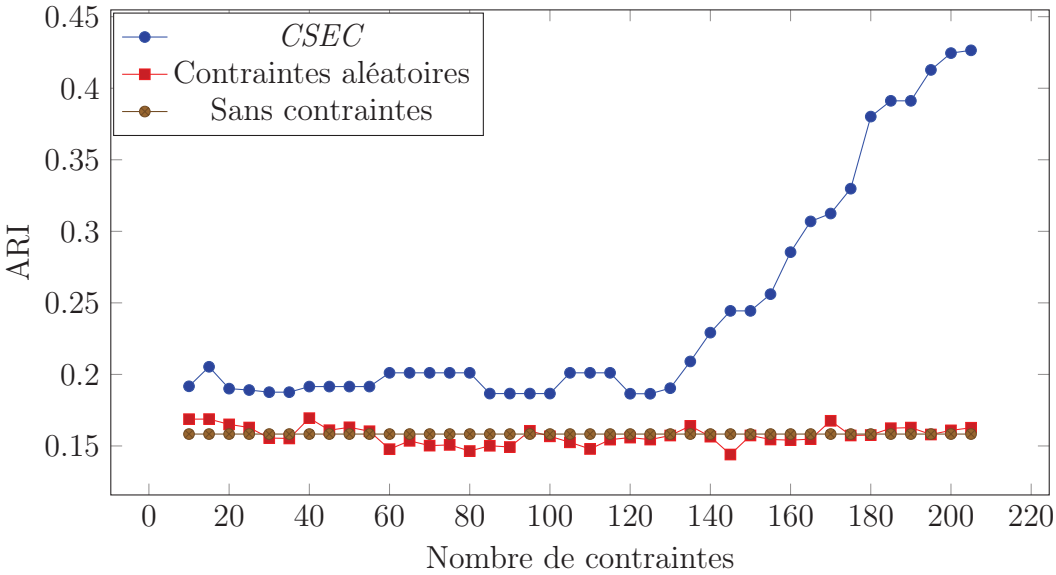


FIGURE 6.11 – Les résultats ARI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur les trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en utilisant les bandes spectrales combinées aux indices radiométriques.

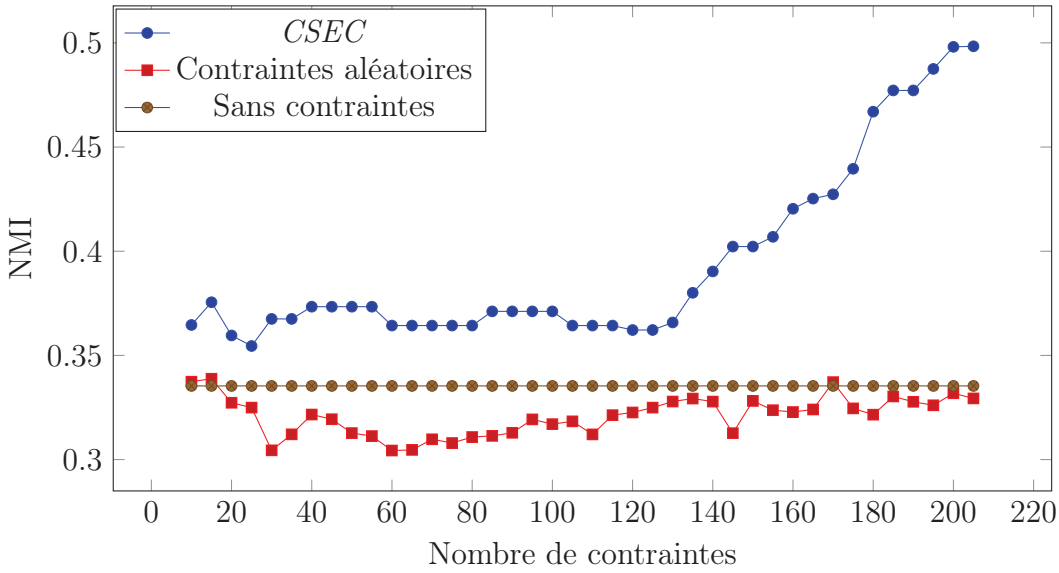


FIGURE 6.12 – Le résultats NMI obtenus par les méthodes : (i) clustering sans contraintes (ii) clustering avec contraintes aléatoires et (iii) *CSEC* sur les trois sites d'étude : la Basse Plaine de l'Aude, la MMCA et le Pic Saint Loup, en utilisant les bandes spectrales combinées aux indices radiométriques.

6.5 Conclusion

Nous avons décrit une méthode d'analyse semi-supervisée de séries temporelles d'images satellites. Cette méthode se base sur du clustering par contraintes **Must-link** et **Cannot-link**. Les séries temporelles sont segmentées puis analysées afin d'identifier des entités spatio-temporelles d'intérêt. L'évolution de ces entités spatio-temporelles est ensuite illustrée en utilisant des graphes d'évolution et des synopsis tels que décrit dans les Chapitres 4 et 5. Notre objectif est d'identifier des groupes homogènes d'entités spatio-temporelles évoluant similairement. Afin d'obtenir un partitionnement d'entités spatio-temporelles correspondant aux besoins et aux attentes de l'utilisateur (expert), nous avons utilisé un algorithme de clustering par contraintes MPCKmeans. Cet algorithme utilise les connaissances expertes pour guider le processus de clustering. Les connaissance expertes sont exprimées sous forme de contraintes **Must-link** et **Cannot-link**.

Nous avons introduit dans ce chapitre une méthode de sélection de contraintes nommée *CSEC*. Cette méthode se base sur plusieurs résultats de clustering. Les entités spatio-temporelles sont analysées par différents algorithmes de clustering afin de générer plusieurs partitionnements. Les différents partitionnements sont ensuite inspectés afin d'identifier les paires d'entités susceptibles d'être mal classées. Ces paires d'entités spatio-temporelles sont utilisées afin de guider le processus de clustering.

Nous avons validé cette méthode sur quatre sites d'étude : la Basse Plaine de l'Aude, la Vallée du Libron , la MMCA et le Pic Saint Loup, en utilisant des séries

temporelles Landsat et Spot.

Les résultats obtenus montre que le clustering semi-supervisé des STIS obtient de meilleurs performances que le clustering non supervisé. En effet, l'intégration de contraintes a permet de guider le processus de clustering afin de générer des clusters cohérents. Les résultats soulignent aussi l'importance du choix des contraintes. La méthode *CSEC* permet de sélectionner des contraintes profitables à l'algorithme de clustering en comparaison à la sélection aléatoire des contraintes. Les contraintes aléatoires peuvent dégrader les performances de clustering au lieu de les améliorer.

Afin d'évaluer notre méthode d'une manière exhaustive, nous proposons en perspective de comparer la méthode *CSEC* aux méthodes de sélection de contraintes proposées dans la littérateur.

Chapitre 7

Conclusion et Perspectives

Sommaire

7.1 Synthèse des travaux

L'objectif de notre thèse est l'analyse des Séries Temporelles d'Images Satellites (STIS) pour le suivi de l'évolution des habitats naturels est semi-naturels. Afin d'analyser ces STIS, nous avons utilisé des méthodes orientées objet. Ces méthodes exploitent des images satellites segmentées. Nous avons d'abord identifié les entités spatio-temporelles d'intérêt parmi les objets de la série. Puis, nous avons illustré leur évolution en utilisant des graphes d'évolution. Ces graphes permettent de représenter les trois dimensions spatiale, temporelle et spectrale des séries temporelles d'images satellites.

Nous avons utilisé les graphes pour identifier, mettre en évidence et organiser les différents types d'évolution dans les STIS. Nous avons d'abord exploité l'information portée par les graphes, afin d'estimer la distance entre les profils d'évolution des entités. Ensuite, nous avons utilisé des méthodes d'apprentissage automatique non supervisé et semi-supervisé pour regrouper les profils d'évolution similaires.

Nous avons proposé des méthodes qui permettent d'explorer les différentes représentations des images satellites. Nous avons comparé d'une part, la représentation pixel à la représentation objet, et d'autre part, l'analyse supervisée à l'analyse semi-supervisée. Afin d'évaluer et valider ces contributions, nous avons mené des expérimentations sur des séries temporelles d'images satellites décrivant quatre sites d'étude.

7.2 Contributions

Dans nos travaux de thèse, nous avons proposé trois contributions. Dans la première contribution, nous avons abordé l'analyse de séries temporelles annuelles. Cette approche a été évaluée sur deux sites d'étude. Chaque site d'étude est décrit par une série temporelle composée de six images Landsat-5, acquises sur une période de huit mois. L'évolution de chaque site a été illustrée au niveau objet en utilisant les graphes d'évolution et au niveau pixel. Puis, ces évolutions ont été regroupées.

Les résultats ont été analysés quantitativement et qualitativement. L'analyse quantitative a permis de comparer le regroupement automatique à un regroupement expert. Nous avons utilisé des indices de validation externes pour calculer la similarité entre les deux regroupements. Les résultats obtenus ont montré que la représentation par objets est pertinente pour la description des STIS. Nous avons aussi effectué une analyse qualitative en observant les différents clusters résultats, nous avons ainsi identifié les patrons d'évolution mis en évidence.

La deuxième contribution a permis d'effectuer une analyse pluriannuelle et multi-site. Nous avons mené cette analyse sur trois sites d'étude. Chaque site est décrit par une série temporelle d'images satellites pluriannuelle. Dans un premier temps, nous avons identifié les entités spatio-temporelles d'intérêt de chaque site, puis nous avons illustré leur évolution en utilisant des graphes d'évolution. Nous avons considéré uniquement les images satellites ne contenant pas d'information redondante. Dans un

deuxième temps, nous avons regroupé les graphes d'évolution des trois sites simultanément, puis nous avons calculé la distance entre chaque paire. Les différentes séries temporelles sont caractérisées par un échantillonnage irrégulier et sont de tailles différentes. Afin de pallier ce problème, nous avons utilisé la mesure Dynamique Time Warpping (DTW). De même que pour la première contribution, les résultats ont été évalué quantitativement et qualitativement. Les résultats ont montré que notre méthode est capable de mettre en évidence des entités spatio-temporelles similaires dans chaque site mais aussi des entités similaires dans les trois sites. La mesure DTW a permis d'estimer pertinemment la distance entre les profils des entités. L'analyse quantitative et l'inspection des clusters a permis d'observer les profils types identifiés.

Les deux premières contributions ont permis de regrouper et de mettre en évidence les évolutions de différents sites. Cependant certains types d'évolution ont été mal regroupés, nous avons par exemple identifié une confusion entre l'évolution des différents types de couverts végétaux.

Notre dernière contribution introduit un système d'analyse basé sur l'apprentissage semi-supervisé. Notre objectif est d'une part d'améliorer les performances du partitionnement en utilisant des données étiquetées. D'autre part, ces données sont utilisées pour guider le processus de clustering afin de l'adapter aux besoins d'un utilisateur et une tache spécifique. L'intérêt du clustering semi-supervisé est d'utiliser le minimum suffisant de données étiquetées. Le choix des données à étiqueter est une étape critique. Nous avons proposé une méthode de sélection basée sur le clustering par ensemble, qui permet d'associer à chaque paire d'entités un score de confiance. La méthode identifie ensuite les paires ayant les scores minimaux. Ces paires d'entités sont les plus susceptibles d'être mal regroupées, elles sont ainsi sélectionnées pour être étiquetées. Nous avons évalué cette méthode en utilisant les entités spatio-temporelles et les profils construits dans la première et la deuxième contribution. Les résultats ont montré que les données sélectionnées ont permis d'améliorer les performances du partitionnement sur les quatre sites d'étude, en utilisant les profils annuels ainsi que les profils pluriannuels.

7.3 Perspectives

Nous travaux de thèse ont permis d'explorer conjointement le domaine de l'apprentissage automatique et de la télédétection. Nous avons exploité les séries temporelles d'images satellites pour le suivi de l'évolution des habitats naturels et semi-naturels.

Nous avons introduit une méthode qui permet d'analyser des images satellites segmentées. D'abord, nous identifions dans les images les objets d'intérêt à analyser. Puis, nous illustrons leur évolution en utilisant des graphes. Enfin, nous appliquons un algorithme de clustering sur les graphes pour identifier des évolutions similaires. Dans un premier temps, nous proposons des perspectives pour chacune de ces étapes

afin d'approfondir notre analyse.

— Nous avons introduit une méthode d'analyse objet. Nous traitons des séries temporelles d'images satellites segmentées. La segmentation fait partie de l'étape de pré-traitement des données. Elle permet d'identifier les objets qui seront analysés. Ce pré-traitement impacte d'une manière indirecte notre méthode d'analyse. Une sous-segmentation ou une sur-segmentation permet d'identifier des objets différents et ainsi d'obtenir des résultats différents. Des expérimentations en utilisant différentes granularités de segmentation permettraient d'observer l'impact de la segmentation sur les résultats de notre analyse.

— Notre méthode permet d'illustrer l'évolution des entités spatio-temporelles en utilisant des graphes. Nous avons utilisé les objets résultats de la segmentation pour construire ces graphes. Pour analyser ces graphes, nous avons principalement exploité l'information radiométrique des objets. Nous avons aussi exploité l'information géométrique des objets en prenant en compte leur taille. En ce qui concerne la structure du graphe, nous avons exploité l'ordre des objets dans les graphes (les graphes construits sont orientés). La structure des graphes peut être un indicateur caractéristique des évolutions des entités spatio-temporelles. Le regroupement des graphes peut être effectué en se basant sur leur structure. Ainsi, nous proposons d'utiliser des algorithmes de graphe *matching*. Ces algorithmes permettent de comparer les nœuds et les arcs de deux graphes pour identifier des sous-graphes communs.

— En ce qui concerne l'étape de clustering, plusieurs algorithmes de clustering existent. Nous avons exploité trois algorithmes de clustering : spectral, hiérarchique agglomératif et Kmeans. Nous avons observé que les résultats sont impactés par l'algorithme de clustering utilisé. En outre, les méthodes proposées dans la littérature utilisent principalement l'algorithme Kmeans. Des expérimentations génériques sont nécessaires, afin observer les performances des algorithmes de clustering pour l'analyse des STIS en utilisant les descripteurs radiométriques.

— En ce qui concerne le clustering par contraintes, nous avons exploité l'algorithme MPCKmeans, Cette algorithme est une variante du Kmeans, Il permet de regrouper les entités en se basant sur la mesure de distance euclidienne. Nous proposons de l'adapter pour gérer des séries temporelles de différentes tailles en lui intégrant la mesure dynamic time warpping.

Dans un deuxième temps, nous proposons de nouvelles contributions qui s'alignent avec nos travaux actuels.

— Dans notre dernière contribution, nous utilisons des données étiquetées (contraintes) pour guider le processus de clustering des séries temporelles. Nous proposons une méthode qui sélectionne les contraintes au début du processus. Ces contraintes sont étiquetées par un expert. Elles sont par la suite

exploitées pendant le clustering. Il existe cependant des méthodes permettant de combiner l'étape de sélection et l'étape de clustering. Ces méthodes sont connues sous le nom d'apprentissage actif. Le processus de clustering actif est illustré dans la Figure 7.1.

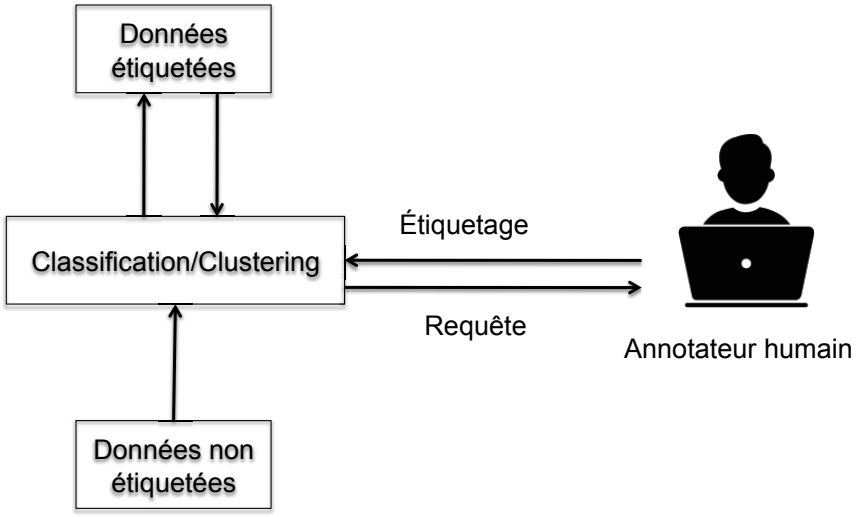


FIGURE 7.1 – Processus d'apprentissage actif.

Au cours du clustering actif, l'utilisateur interagit avec l'algorithme d'apprentissage d'une manière itérative. À chaque itération, l'algorithme propose à l'utilisateur d'annoter des contraintes pour les utiliser. Chaque itération permet de modifier le regroupement, le clustering actif permet ainsi de réduire l'ensemble de données étiquetées nécessaires.

En outre, cette méthode implique la présence de l'expert tout le long du processus alors qu'il intervient uniquement sur une partie. Afin de répondre à cette problématique, des travaux récents (CHAHDI, 2017) ont proposés d'automatiser entièrement le processus d'annotation en utilisant des ressources externes telles que les ontologies. Le principe de ces travaux est de requêter ces ressources au lieu d'un expert. Ces ressources formalisent, stockent et organisent les connaissances expertes. Cette approche permet à l'utilisateur d'intervenir uniquement pendant la construction de la ressource de connaissances, tandis que le processus de clustering devient entièrement automatique.

— Dans les trois contributions présentées dans cette thèse, nous avons analysé plusieurs séries temporelles afin d'identifier des profils d'évolution similaires. Cependant, considérer les séries temporelles d'une façon individuelle permettra d'identifier des connaissances utiles pour le suivi de l'évolution des habitats. Nous nous intéressons plus particulièrement aux phénomènes récurrents. Nous proposons d'analyser les graphes d'évolutions indépendamment

pour identifier des évolutions saisonnières. Nous proposons d'étiqueter les objets du graphe afin de construire des séquences de motifs. Puis, d'exploiter des méthodes d'analyse de motifs séquentiels, afin d'identifier des sous-séquences fréquentes (Inokuchi et Washio, 2012 ; Vo et collab., 2017).

Bibliographie

Abin, A. A. 2016, «Querying Beneficial Constraints Before Clustering Using Facility Location Analysis», , p. 1–12.

Abin, A. A. et H. Beigy. 2014, «Active selection of clustering constraints : A sequential approach», *Pattern Recognition*, vol. 47, n° 3, doi :10.1016/j.patcog.2013.09.034, p. 1443–1458, ISSN 00313203. URL `http://dx.doi.org/10.1016/j.patcog.2013.09.034`.

Abin, A. A. et H. Beigy. 2015, «Active constrained fuzzy clustering : A multiple kernels learning approach», *Pattern Recognition*, vol. 48, n° 3, doi :10.1016/j.patcog.2014.09.008, p. 953–967, ISSN 00313203.

Arbelaitz, O., I. Gurrutxaga, J. Muguerza, J. M. Pérez et I. Perona. 2013, «An extensive comparative study of cluster validity indices», *Pattern Recognition*, vol. 46, n° 1, p. 243–256.

Arthur, D. et S. Vassilvitskii. 2007, «k-means++ : The advantages of careful seeding», dans *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, p. 1027–1035.

Baatz, M. et A. Schäpe. 2000, «Multiresolution segmentation : an optimization approach for high quality multi-scale image segmentation», .

Basu, S., A. Banerjee et R. J. Mooney. 2004, «Active semi-supervision for pairwise constrained clustering», dans *Proceedings of the 2004 SIAM international conference on data mining*, SIAM, p. 333–344.

Basu, S., I. Davidson et K. Wagstaff. 2008, *Constrained clustering : Advances in algorithms, theory, and applications*, CRC Press.

Bilenko, M., S. Basu et R. J. Mooney. 2004, «Integrating constraints and metric learning in semi-supervised clustering», dans *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, ACM, New York,

NY, USA, ISBN 1-58113-838-5, p. 11–, doi :10.1145/1015330.1015360. URL http://doi.acm.org/10.1145/1015330.1015360.

Blaschke, T. 2010, «Object based image analysis for remote sensing», *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, n° 1, doi :https://doi.org/10.1016/j.isprsjprs.2009.06.004, p. 2 – 16, ISSN 0924-2716. URL http://www.sciencedirect.com/science/article/pii/S0924271609000884.

Bontemps, S., P. Bogaert, N. Titeux et P. Defourny. 2008, «An object-based change detection method accounting for temporal dependences in time series with medium to coarse spatial resolution», *Remote Sensing of Environment*, vol. 112, n° 6, doi :https://doi.org/10.1016/j.rse.2008.03.013, p. 3181 – 3191, ISSN 0034-4257. URL http://www.sciencedirect.com/science/article/pii/S0034425708001119.

Caliński, T. et J. Harabasz. 1974, «A dendrite method for cluster analysis», *Communications in Statistics-theory and Methods*, vol. 3, n° 1, p. 1–27.

Chahdi, H. 2017, *Apports des ontologies à l'analyse exploratoire des images satellitaires*, thèse de doctorat, Université montpellier II.

Chahdi, H., N. Grozavu, I. Mougenot, L. Berti-Équille et Y. Bennani. 2016, «On the use of ontology as A priori knowledge into constrained clustering», dans *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*, p. 632–641, doi :10.1109/DSAA.2016.72. URL https://doi.org/10.1109/DSAA.2016.72.

Cohen-Addad, V., V. Kanade, F. Mallmann-Trenn et C. Mathieu. 2018, «Hierarchical clustering : Objective functions and algorithms», dans *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, p. 378–397.

Comaniciu, D. et P. Meer. 2002, «Mean shift : A robust approach toward feature space analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, p. 603–619.

Conchedda, G., L. Durieux et P. Mayaux. 2008, «An object-based method for mapping and change analysis in mangrove ecosystems», *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, n° 5, p. 578–589.

Cortes, C. et V. Vapnik. 1995, «Support-vector networks», *Machine learning*, vol. 20, n° 3, p. 273–297.

Davidson, I. et S. Basu. 2007, «A survey of clustering with instance level», *Constraints*, vol. 1, p. 2.

Davidson, I. et S. Ravi. 2005a, «Clustering with constraints : Feasibility issues and the k-means algorithm», dans *Proceedings of the 2005 SIAM international conference on data mining*, SIAM, p. 138–149.

Davidson, I. et S. S. Ravi. 2005b, «Agglomerative hierarchical clustering with constraints : Theoretical and empirical results», dans *Knowledge Discovery in Databases : PKDD 2005*, édité par A. M. Jorge, L. Torgo, P. Brazdil, R. Camacho et J. Gama, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 59–70.

Davies, D. L. et D. W. Bouldin. 1979, «A cluster separation measure», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, n° 2, doi :10.1109/TPAMI.1979.4766909, p. 224–227, ISSN 0162-8828.

Desclée, B., P. Bogaert et P. Defourny. 2006, «Forest change detection by statistical object-based method», *Remote Sensing of Environment*, vol. 102, n° 1-2, p. 1–11.

Dey, V., Y. Zhang et M. Zhong. 2010, *A review on image segmentation techniques with remote sensing perspective*, na.

Ding, H., G. Trajcevski, P. Scheuermann, X. Wang et E. Keogh. 2008, «Querying and mining of time series data : Experimental comparison of representations and distance measures», *Proc. VLDB Endow.*, vol. 1, n° 2, doi : 10.14778/1454159.1454226, p. 1542–1552, ISSN 2150-8097. URL `http://dx.doi.org/10.14778/1454159.1454226`.

Escadafal, R. 1993, «Remote sensing of soil color : Principles and applications», *Remote Sensing Reviews*, vol. 7, n° 3-4, doi :10.1080/02757259309532181, p. 261–279. URL `https://doi.org/10.1080/02757259309532181`.

Ferri, C., J. Hernández-Orallo et R. Modroiu. 2009, «An experimental comparison of performance measures for classification», *Pattern Recognition Letters*, vol. 30, n° 1, p. 27–38.

Filippone, M., F. Camastra, F. Masulli et S. Rovetta. 2008, «A survey of kernel and spectral methods for clustering», *Pattern recognition*, vol. 41, n° 1, p. 176–190.

Frey, B. J. et D. Dueck. 2007, «Clustering by passing messages between data points», *science*, vol. 315, n° 5814, p. 972–976.

Gan, G., C. Ma et J. Wu. 2007, *Data clustering : theory, algorithms, and applications*, vol. 20, Siam.

Gao, B.-C. 1996, «Ndwi : A normalized difference water index for remote sensing of vegetation liquid water from space», *Remote sensing of environment*, vol. 58, n° 3, p. 257–266.

Girard, M.-C. et C.-M. Girard. 2010, *Traitement des données de télédétection-2e éd. : Environnement et ressources naturelles*, Dunod.

Gonçalves, R., J. Zullo, B. F. d. Amaral, P. P. Coltri, E. P. M. d. Sousa et L. A. S. Romani. 2014, «Land use temporal analysis through clustering techniques on satellite image time series», dans *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, IEEE, p. 2173–2176.

Grira, N., M. Crucianu et N. Boujemaa. 2006, «Fuzzy clustering with pairwise constraints for knowledge-driven image categorisation», *IEE Proceedings - Vision, Image and Signal Processing*, vol. 153, n° 3, doi :10.1049/ip-vis:20050060, p. 299–304, ISSN 1350-245X.

Guttler, F., D. Ienco, J. Nin, M. Teisseire et P. Poncelet. 2017, «A graph-based approach to detect spatiotemporal dynamics in satellite image time series», *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, p. 92–107.

Guyet, T. et H. Nicolas. 2016, «Long term analysis of time series of satellite images», *Pattern Recogn. Lett.*, vol. 70, n° C, doi :10.1016/j.patrec.2015.11.005, p. 17–23, ISSN 0167-8655. URL `http://dx.doi.org/10.1016/j.patrec.2015.11.005`.

Hämäläinen, J., S. Jauhiainen et T. Kärkkäinen. 2017, «Comparison of internal clustering validation indices for prototype-based clustering», *Algorithms*, vol. 10, n° 3, p. 105.

Hansen, L. K. et P. Salamon. 1990, «Neural network ensembles», *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, n° 10, p. 993–1001.

Herbert, S. 1983, «Administration et processus de décision», *Economica*.

Huete, A. 1988, «A soil-adjusted vegetation index (savi)», *Remote Sensing of Environment*, vol. 25, n° 3, doi :https://doi.org/10.1016/0034-4257(88)90106-X, p. 295 – 309, ISSN 0034-4257. URL `http://www.sciencedirect.com/science/article/pii/003442578890106X`.

Image, S. 1988, «Spot user's handbook», *Centre National d'Etude Spatiale (CNES) and SPOT Image*, vol. 1, p. 3.

Inokuchi, A. et T. Washio. 2012, «Frissminer : mining frequent graph sequence patterns induced by vertices», *IEICE TRANSACTIONS on Information and Systems*, vol. 95, n° 6, p. 1590–1602.

Jiang, Z., S. Shekhar, X. Zhou, J. K. Knight et J. Corcoran. 2013, «Focal-test-based spatial decision tree learning : A summary of results», dans *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10,*

2013, p. 320–329, doi :10.1109/ICDM.2013.96. URL https://doi.org/10.1109/ICDM.2013.96.

Julea, A., N. Méger, P. Bolon, C. Rigotti, M.-P. Doin, C. Lasserre, E. Trouvé et V. N. Lazarescu. 2011, «Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns», *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, n° 4, p. 1417–1430.

Kaufman, L. et P. Rousseeuw. 1987, *Clustering by means of medoids*, North-Holland.

Kaufman, L. et P. J. Rousseeuw. 2009, *Finding groups in data : an introduction to cluster analysis*, vol. 344, John Wiley & Sons.

Khiali, L., D. Ienco et M. Teisseire. 2018, «Object-oriented satellite image time series analysis using a graph-based representation», *Ecological Informatics*, vol. 43, doi :https://doi.org/10.1016/j.ecoinf.2017.11.003, p. 52 – 64, ISSN 1574-9541. URL http://www.sciencedirect.com/science/article/pii/S1574954117301851.

Khiali, L., M. Ndiath, S. Alleaume, D. Ienco, K. Ose et M. Teisseire. 2019, «Detection of spatio-temporal evolutions on multi-annual satellite image time series : A clustering based approach», *International Journal of Applied Earth Observation and Geoinformation*, vol. 74, doi :https://doi.org/10.1016/j.jag.2018.07.014, p. 103 – 119, ISSN 0303-2434. URL http://www.sciencedirect.com/science/article/pii/S0303243418304781.

Kurtz, C., N. Passat, P. Gançarski et A. Puissant. 2010, «Multi-resolution region-based clustering for urban analysis», *International Journal of Remote Sensing*, vol. 31, n° 22, doi :10.1080/01431161.2010.512312, p. 5941–5973. URL http://dx.doi.org/10.1080/01431161.2010.512312.

Laliberte, A. S., A. Rango, K. M. Havstad, J. F. Paris, R. F. Beck, R. McNeely et A. L. Gonzalez. 2004, «Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern new mexico», *Remote Sensing of Environment*, vol. 93, n° 1-2, p. 198–210.

Lelis, L. et J. Sander. 2009, «Semi-supervised density-based clustering», dans *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, IEEE Computer Society, Washington, DC, USA, ISBN 978-0-7695-3895-2, p. 842–847, doi :10.1109/ICDM.2009.143. URL https://doi.org/10.1109/ICDM.2009.143.

MacQueen, J. et collab.. 1967, «Some methods for classification and analysis of multivariate observations», dans *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, Oakland, CA, USA, p. 281–297.

Masseglia, F., M. Teisseire et P. Poncelet. 2004, «Extraction de motifs séquentiels», *Problemes et methodes*.

Mathieu, R., C. Freeman et J. Aryal. 2007, «Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery», *Landscape and Urban Planning*, vol. 81, n° 3, p. 179–192.

Maulik, U. et D. Chakraborty. 2011, «A self-trained ensemble with semisupervised svm : An application to pixel classification of remote sensing imagery», *Pattern Recognition*, vol. 44, n° 3, p. 615–623.

Müller, M. 2007, *Information retrieval for music and motion*, vol. 2, Springer, 69-84 p..

Nasrabadi, N. M. 2007, «Pattern recognition and machine learning», *Journal of electronic imaging*, vol. 16, n° 4, p. 049 901.

Petitjean, F., J. Inglada et P. Gançarski. 2012a, «Satellite image time series analysis under time warping», *IEEE Trans. Geoscience and Remote Sensing*, vol. 50, n° 8, doi :10.1109/TGRS.2011.2179050, p. 3081–3095. URL `https://doi.org/10.1109/TGRS.2011.2179050`.

Petitjean, F., J. Inglada et P. Gançarskv. 2011, «Clustering of satellite image time series under time warping», dans *Analysis of Multi-temporal Remote Sensing Images (Multi-Temp), 2011 6th International Workshop on the*, IEEE, p. 69–72.

Petitjean, F., C. Kurtz, N. Passat et P. Gançarski. 2012b, «Spatio-temporal reasoning for the classification of satellite image time series», *Pattern Recognition Letters*, vol. 33, n° 13, p. 1805–1815.

Qin, Y., Z. Niu, F. Chen, B. Li et Y. Ban. 2013, «Object-based land cover change detection for cross-sensor images», *International Journal of Remote Sensing*, vol. 34, n° 19, doi :10.1080/01431161.2013.805282, p. 6723–6737. URL `http://dx.doi.org/10.1080/01431161.2013.805282`.

Rai, P. et S. Singh. 2010, «A survey of clustering techniques», *International Journal of Computer Applications*, vol. 7, n° 12, p. 1–5.

Rendón, E., I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz et H. E. Arzate. 2011, «A comparison of internal and external cluster validation indexes», dans *Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, AMERICAN-MATH'11/CEA'11, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, ISBN 978-960-474-270-7, p. 158–163. URL `http://dl.acm.org/citation.cfm?id=1959666.1959695`.

Romani, L., R. Gonçalves, B. Amaral, D. Chino, J. Zullo, C. Traina, E. Sousa et A. Traina. 2011, «Clustering analysis applied to ndvi/noaa multitemporal images to improve the monitoring process of sugarcane crops», dans *Analysis of Multi-temporal Remote Sensing Images (Multi-Temp), 2011 6th International Workshop on the*, IEEE, p. 33–36.

Rouse Jr, J., R. Haas, J. Schell et D. Deering. 1974, «Monitoring vegetation systems in the great plains with erts», .

Rousseeuw, P. J. 1987, «Silhouettes : A graphical aid to the interpretation and validation of cluster analysis», *Journal of Computational and Applied Mathematics*, vol. 20, doi :https://doi.org/10.1016/0377-0427(87)90125-7, p. 53 – 65, ISSN 0377-0427. URL `http://www.sciencedirect.com/science/article/pii/0377042787901257`.

Roy, M., S. Ghosh et A. Ghosh. 2014, «A novel approach for change detection of remotely sensed images using semi-supervised multiple classifier system», *Information Sciences*, vol. 269, p. 35–47.

Ruiz, C., M. Spiliopoulou et E. Menasalvas. 2007, «C-dbscan : Density-based clustering with constraints», dans *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, édité par A. An, J. Stefanowski, S. Ramanna, C. J. Butz, W. Pedrycz et G. Wang, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 216–223.

Sakoe, H. et S. Chiba. 1978, «Dynamic programming algorithm optimization for spoken word recognition», *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, n° 1, p. 43–49.

Salzberg, S. L. 1994, «C4. 5 : Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993», *Machine Learning*, vol. 16, n° 3, p. 235–240.

Schölkopf, B., A. J. Smola et collab.. 2002, *Learning with kernels : support vector machines, regularization, optimization, and beyond*, MIT press.

Schuster, C., T. Schmidt, C. Conrad, B. Kleinschmit et M. Förster. 2015, «Grassland habitat mapping by intra-annual time series analysis–comparison of rapideye and terrasar-x satellite data», *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, p. 25–34.

Sentinel, E. 2013, «User handbook», .

Settles, B. 2010, «Active Learning Literature Survey», .

Sun, S. 2013, «A survey of multi-view machine learning», *Neural Computing and Applications*, vol. 23, n° 7-8, p. 2031–2038.

Tan, K., J. Zhu, Q. Du, L. Wu et P. Du. 2016, «A novel tri-training technique for semi-supervised classification of hyperspectral images based on diversity measurement», *Remote Sensing*, vol. 8, n° 9, doi :10.3390/rs8090749, p. 749. URL https://doi.org/10.3390/rs8090749.

Tan, P.-N., M. Steinbach et V. Kumar. 2005, *Introduction to Data Mining, (First Edition)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, ISBN 0321321367.

Thomas, M. 1991, «Cover and joy a. thomas : Elements of information theory», *Wiley*, vol. 4, p. 10.

Van Craenendonck, T. et H. Blockeel. 2017, «Constraint-based clustering selection», *Machine Learning*, vol. 106, n° 9-10, p. 1497–1521.

Vapnik, V. 1998, *Statistical learning theory. 1998*, vol. 3, Wiley, New York.

Vega-Pons, S. et J. Ruiz-Shulcloper. 2011, «A survey of clustering ensemble algorithms», *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, n° 03, p. 337–372.

Vo, B., S. Pham, T. Le et Z.-H. Deng. 2017, «A novel approach for mining maximal frequent patterns», *Expert Systems with Applications*, vol. 73, p. 178–186.

Von Luxburg, U. 2007, «A tutorial on spectral clustering», *Statistics and computing*, vol. 17, n° 4, p. 395–416.

Vu, N., Viet Vu et Labroche et B. Bouchon-Meunier. 2009, «Leader Ant Clustering with Constraints», dans *IEEE RIVF International Conference on Computing and Telecommunication Technologies*, IEEE, Da Nang, Vietnam, p. 79–86, doi :10.1109/RIVF.2009.5174648. URL https://hal.archives-ouvertes.fr/hal-01297949.

Vu, V. V., N. Labroche et B. Bouchon-Meunier. 2012, «Improving constrained clustering with active query selection», *Pattern Recognition*, vol. 45, n° 4, doi :10.1016/j.patcog.2011.10.016, p. 1749–1758, ISSN 00313203.

Wagstaff, K. et C. Cardie. 2000, «Clustering with instance-level constraints», dans *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-707-2, p. 1103–1110. URL http://dl.acm.org/citation.cfm?id=645529.658275.

Wagstaff, K., C. Cardie, S. Rogers et S. Schrödl. 2001, «Constrained k-means clustering with background knowledge», dans *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, Morgan Kaufmann

Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-778-1, p. 577–584. URL http://dl.acm.org/citation.cfm?id=645530.655669.

Wagstaff, K. L. 2006, «Value, cost, and sharing : Open issues in constrained clustering», dans *International workshop on knowledge discovery in inductive databases*, Springer, p. 1–10.

Wang, X. et I. Davidson. 2010, «Flexible constrained spectral clustering», dans *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, ACM, New York, NY, USA, ISBN 978-1-4503-0055-1, p. 563–572, doi :10.1145/1835804.1835877. URL http://doi.acm.org/10.1145/1835804.1835877.

Wemmert, C., A. Puissant, G. Forestier et P. Gançarski. 2009, «Multiresolution remote sensing image clustering», *IEEE Geosci. Remote Sensing Lett.*, vol. 6, n° 3, doi :10.1109/LGRS.2009.2020825, p. 533–537. URL https://doi.org/10.1109/LGRS.2009.2020825.

Xu, D. et Y. Tian. 2015, «A comprehensive survey of clustering algorithms», *Annals of Data Science*, vol. 2, n° 2, doi :10.1007/s40745-015-0040-1, p. 165–193, ISSN 2198-5812. URL https://doi.org/10.1007/s40745-015-0040-1.

Xu, R. et D. Wunsch. 2005, «Survey of clustering algorithms», *IEEE Transactions on Neural Networks*, vol. 16, n° 3, doi :10.1109/TNN.2005.845141, p. 645–678, ISSN 1045-9227.

Yarowsky, D. 1995, «Unsupervised word sense disambiguation rivaling supervised methods», dans *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 189–196.

Zhang, N., Y. Hong, Q. Qin et L. Liu. 2013, «Vsdi : a visible and shortwave infrared drought index for monitoring soil and vegetation moisture based on optical remote sensing», *International journal of remote sensing*, vol. 34, n° 13, p. 4585–4609.

Zhang, Z., P. Tang et T. Corpetti. 2016, «Satellite image time series clustering via affinity propagation», dans *2016 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016, Beijing, China, July 10-15, 2016*, p. 2419–2422, doi :10.1109/IGARSS.2016.7729624. URL https://doi.org/10.1109/IGARSS.2016.7729624.

Zhou, Z.-H. 2012, *Ensemble methods : foundations and algorithms*, Chapman and Hall/CRC.

Zhu, L., P. Xiao, X. Feng, X. Zhang, Y. Huang et C. Li. 2016, «A co-training, mutual learning approach towards mapping snow cover from multi-temporal high-spatial resolution satellite imagery», *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, p. 179–191.

Abstract

Nowadays, remotely sensed images constitute a rich source of information that can be leveraged to support several applications including risk prevention, land use planning, land cover classification and many other several tasks. In this thesis, Satellite Image Time Series (SITS) are analysed to depict the dynamic of natural and semi-natural habitats. The objective is to identify, organize and highlight the evolution patterns of these areas.

We introduce an object-oriented method to analyse SITS that consider segmented satellites images. Firstly, we identify the evolution profiles of the objects in the time series. Then, we analyse these profiles using machine learning methods. To identify the evolution profiles, we explore all the objects to select a subset of objects (spatio-temporal entities/reference objects) to be tracked. The evolution of the selected spatio-temporal entities is described using evolution graphs.

To analyse these evolution graphs, we introduced three contributions. The first contribution explores annual SITS. It analyses the evolution graphs using clustering algorithms, to identify similar evolutions among the spatio-temporal entities. In the second contribution, we perform a multi-annual cross-site analysis. We consider several study areas described by multi-annual SITS. We use the clustering algorithms to identify intra and inter-site similarities. In the third contribution, we introduce à semi-supervised method based on constrained clustering. We propose a method to select the constraints that will be used to guide the clustering and adapt the results to the user needs.

Our contributions were evaluated on several study areas. The experimental results allow to pinpoint relevant landscape evolutions in each study sites. We also identify the common evolutions among the different sites. In addition, the constraint selection method proposed in the constrained clustering allows to identify relevant entities. Thus, the results obtained using the unsupervised learning were improved and adapted to meet the user needs.