Découverte de motifs d'évolution significatifs dans les séries temporelles d'images satellites

François Petitjean, Florent Masseglia, Pierre Gancarski

Découverte de motifs d'évolution significatifs dans les séries temporelles d'images satellites

Francois Petitjean*[1], Florent Masseglia**, Pierre Gancarski*

*LSIIT (UMR 7005 CNRS/UdS) – Bd Sébastien Brant – 67412 Illkirch – France
{fpetitjean,gancarski}@unistra.fr
**INRIA – 2004, route des Lucioles – BP 93 – 06902 Sophia Antipolis – France
florent.masseglia@inria.fr

Résumé. Les séries temporelles d'images satellites (ou Satellite Image Time Series – SITS) sont d'importantes sources d'informations sur l'évolution du territoire. Étudier ces images permet de comprendre les changements sur des zones précises mais aussi de découvrir des schémas d'évolution à grande échelle. Toutefois, découvrir ces phénomènes impose de répondre à plusieurs défis qui sont liés aux caractéristiques des SITS et à leurs contraintes. Premièrement, chaque pixel d'une image satellite est décrit par plusieurs valeurs (les niveaux radiométriques sur différentes longueurs d'ondes). Deuxièmement, ces motifs d'évolution portent sur des périodes très longues et ne sont pas forcément synchrones selon les régions. Troisièmement, les régions qui ne sont pas concernées par des évolutions signiticatives sont majoritaires et leur domination rend difficile l'extraction des motifs d'évolution. Dans cet article, nous proposons une méthode qui répond à ces difficultés et nous la validons sur une série d'images satellites acquises sur une période de 20 ans.

1 Introduction

La détection du changement est un domaine important de la télédétection et les progrès technologiques récents[2] ont accentué l'attention qui lui est portée. Les séries temporelles d'images satellites (ou Satellite Image Time Series – SITS) sont une source importante d'information pour étudier l'occupation des sols et son évolution. Considérons par exemple la scène illustrée par la figure 1. Chaque image contient quatre régions principales ($R1$ à $R4$) et la série montre leurs évolutions. En juillet 2007, les régions $R1$ et $R2$ étaient principalement constituées d'arbres et sont urbanisées en mai 2009. Ce schéma illustre un phénomène d'urbanisation dans lequel les arbres disparaissent au profit de routes et d'habitations. Ces évolutions constituent une information importante qui peut être utilisée dans diverses applications (Coppin et al., 2004; Campbell, 2007; Jensen, 2007) mais leur découverte soulève deux défis importants.

[1]Ce travail a été partiellement financé par le CNES et Thales Alenia Space.

[2]La prochaine génération de satellites (*e.g.* $Venμs$, *Sentinel-2*) sera capable d'acquérir les images avec une fréquence élevée.

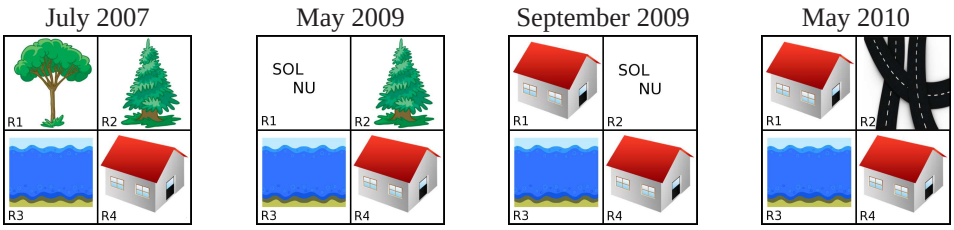


FIG. 1 – *Exemple de changement dans une scène avec le motif "arbres → sol nu → urbain".*

Premièrement, les changements peuvent se produire sur de longues périodes (l'urbanisation peut prendre plusieurs années) ou peuvent être **cycliques** (comme la rotation des cultures). Un exemple de motif intéressant serait "arbres → sol nu → urbain" dans la figure 1 qui est observé pour 50% des pixels (avec un décalage selon les régions).

Deuxièmement, les satellites ne donnent pas de labels automatiquement (comme "arbre" ou "route") aux différentes régions de pixels. Les valeurs des pixels sur chaque bande ne permettent pas de distinguer toutes les surfaces. Il est donc indispensable de considérer toutes les bandes disponibles pour chaque pixel afin d'extraire des motifs d'évolution informatifs.

A notre connaissance, *il n'existe pas de travaux permettant d'extraire des évolutions dans les SITS et montrant des motifs qui s'expriment sur plusieurs bandes*. Les travaux existants étant expérimentalement validés par des motifs de non-évolution limités à une bande. Nous proposons une heuristique d'extraction qui évite l'explosion combinatoire liée aux caractéristiques des motifs d'évolution sur plusieurs bandes dans les SITS. Nous en donnons les limites théoriques et nous montrons des résultats d'expérimentations ainsi qu'un principe de visualisation permettant de synthétiser les résultats.

2 Contexte, motivation et travaux existants

Il existe différentes méthodes pour l'analyse de séries temporelles d'images satellites qui peuvent être réparties en trois catégories. Les méthodes **bi-temporelles** s'intéressent à l'extraction des zones de changement entre deux images, une avant le changement et une après (ex : Bruzzone et Prieto (2000)). Une deuxième famille correspond à des méthodes **statistiques**, s'appliquant à deux images ou plus. Ces méthodes s'intéressent généralement à transformer l'espace des données, comme par une ACP (Howarth et al., 2006). Enfin, une troisième famille de méthode s'intéresse à l'étude de **séries temporelles** d'images satellites. Ces techniques sont généralement dédiées à l'analyse de trajectoires radiométriques de pixels (ex : Kennedy et al. (2007)). Dans le domaine de la télédétection, la nécessité d'utiliser plusieurs bandes est bien connu. En effet, une bande couvre généralement 255 niveaux alors que le nombre de combinaisons sur n bandes pour un pixel peut atteindre 255^n valeurs possibles. Par exemple, une valeur de 200 sur la bande "infra-rouge" peut correspondre à des toits ou à des routes ce qui ne permet pas de les distinguer alors que sur la bande "rouge", les toits et les routes peuvent être différenciés par leurs valeurs (*i.e.* 200 et 280). Cependant, la bande "rouge" ne permet pas de discriminer toutes les surfaces. Par exemple, la réponse spectrale des conifères et des feuillus

dans la bande “rouge” (*i.e.* 120) ne permet pas de les distinguer, mais sur la bande “infra-rouge”, leurs valeurs permettent de les séparer (*i.e. 145 et 200*). Le motif suivant pourrait, par exemple, être découvert grâce à l'utilisation de plusieurs bandes :

$$50\% : (\underbrace{coniferes/resineux}_{Rouge}) \rightarrow (\underbrace{sol}_{Rouge}, \underbrace{sol/coniferes}_{Infra-Rouge}) \rightarrow (\underbrace{toits/route}_{Infra-Rouge})$$

Ce qui pourrait être résumé par un motif plus générique du type “arbres → sol → urbain”, exprimant le fait que dans 50% des pixels de l'image, les arbres sont progressivement remplacés par du sol nu puis par des éléments urbains. Si l'analyse se contente des valeurs mesurées sur la bande infra-rouge, alors seul le motif “sol/coniferes → toits/route” serait découvert, laissant le doute sur le premier élément du motif (sol ou conifères ?). Le problème serait identique avec la bande infra-rouge qui ne permet de découvrir qu'une seule partie du motif. En revanche, avec les deux bandes, l'expert peut rapidement corroborer les informations pour déterminer la texture correspondante. Extraire ce type de motif est un problème proche de celui de l'extraction de motifs séquentiels décrit par Agrawal et Srikant (1995).Toutefois, il ne s'agit pas d'une application directe, dans la mesure où les contraintes et caractéristiques des SITS ne le permettent pas. L'extraction de séquences fréquentes dans les SITS a été introduite par Julea et al. (2006, 2008). Les auteurs y étudient deux types d'applications : la météorologie et l'agronomie. Toutefois, leur proposition décrit des séquences dans des séries d'images ou les valeurs sont mesurées sur une seule bande. Ces travaux sont ensuite étendus dans Julea et al. (2010) avec, en particulier, une fonction attentive aux régions de pixels dans le calcul des motifs.

3 Extraction de motifs séquentiels fréquents

Les motifs séquentiels sont généralement extraits à partir de grands ensembles de données. Ces données contiennent des séquences de valeurs prises dans un ensemble de symboles spécifiques, comme indiqué dans la définition 1 (inspirée des définitions de Agrawal et Srikant (1995)).

Définition 1 *Soit $\mathcal{I} = \{i_1, i_2, ..., i_m\}$, un ensemble de m valeurs (ou* items*). Soit $I = \{t_1, t_2, ...t_n\}$, un sous-ensemble de $\mathcal{I}$. I est un* itemset *noté $(t_1; t_2; ...; t_n)$. Une* séquence *s est un ensemble d'itemsets non vide noté $< s_1, s_2, \ldots, s_n >$ où s_j est un itemset. Une* séquence de données *est une séquence de l'ensemble des données à analyser.*

La définition 2 donne les conditions de l'inclusion entre deux séquences. En d'autres termes, s_1 est incluse dans s_2 si chaque itemset de s_1 est inclus dans un itemset de s_2 en respectant l'ordre dans les séquences (comme illustré par l'exemple 1).

Définition 2 *Soit $s_1 =< a_1, a_2, \ldots, a_n >$ et $s_2 =< b_1, b_2, \ldots, b_m >$ deux séquences. s_1 est* incluse *dans s_2 ($s_1 \prec s_2$) si et seulement si $\exists i_1 < i_2 < \ldots < i_n$ des entiers tels que $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}, \ldots$ $a_n \subseteq b_{i_n}$.*

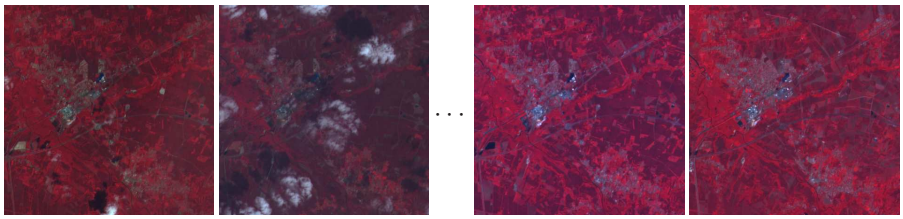


FIG. 2 – *Extrait de la STIS utilisée provenant de la Base de données* LITTORAL – KALIDEOS. *© CNES 2010 – Distribution Spot Image*

Exemple 1 *La séquence* $s_1 =< (3)\ (4;5)\ (8) >$ *est incluse dans la séquence* $s_2 =< (7)\ (3;8)\ (9)\ (4;5;6)\ (8) >$ *(i.e* $s_1 \prec s_2$*) car* $(3) \subseteq (3;8)$, $(4;5) \subseteq (4;5;6)$ *et* $(8) \subseteq (8)$. *Cependant, la séquence* $s_3 =< (3;8;9)\ (4;5) >$ *n'est pas incluse dans* s_2 *car* $(3;8;9)$ *n'est inclus dans aucun itemset de* s_2.

Dans cet article, la caractéristique principale des motifs extraits est leur fréquence. Cette notion est basée sur le nombre d'apparitions d'un motif rapporté au nombre total de séquences de données, comme indiqué par la définition 3. Enfin, pour simplifier la lecture des résultats, seuls les motifs fréquents les plus longs (*i.e.* inclus dans aucun autre motif fréquent) sont conservés.

Définition 3 *Une séquence de données* s_d supporte *une séquence* s *si* $s \prec s_d$. *Soit* D *l'ensemble des séquences de données. Le* support *de* s *dans* D *est donné par le pourcentage de séquences de* D *qui supportent :* $support(s) = |\{s_d \in D | s \prec s_d\}|/|D|$. *Soit* $minSupp$ le support minimum *donné par l'utilisateur, alors une séquence dont le support est supérieur à* $minSupp$ *est dite* fréquente.

4 Prétraitement des images

Les images utilisées dans cet article sont extraites de la base de données Littoral du programme KALIDEOS[3] et se situent dans le sud-ouest de la France. Nous avons extrait de cette base de données une série de 35 images (illustrée en figure 2).

Afin de garantir que les niveaux radiométriques d'un pixel (x, y) soient comparables le long de la série d'images, des corrections ont été effectuées par le CNES avant leur intégration dans la base de données. Ces corrections sont de trois types : des corrections géométriques assurant qu'un pixel (x, y) représente toujours la même zone géographique quelle que soit l'image, des corrections d'éclairement solaire ainsi que des corrections atmosphériques.

Une fois ces corrections effectuées, chaque pixel de la série de 35 images est décrit sur trois bandes : Proche Infra-Rouge (PIR), Rouge (R) et Vert (V). À ces trois bandes, nous ajoutons un quatrième attribut, correspondant à l'index de végétation NDVI couramment utilisé en télédétection et calculé pour un pixel p comme suit : $NDVI = (\text{PIR} - \text{R}) \,/\, (\text{PIR} + \text{R})$.

[3] `http://kalideos.cnes.fr`

Définissons comment les séquences sont construites à partir de cette série d'images. Pour cela, la Définition 4 formalise le concept d'image multivaluée afin de définir la construction des séries temporelles en Définition 5.

Définition 4 *Soit $\mathcal{S}_{image} =< I^1, \ldots, I^{\mathcal{N}} >$ une série de $\mathcal{N}$ images de largeur $\mathcal{W}$ et de hauteur $\mathcal{H}$. Soit $\mathcal{B}$ le nombre de bandes des images. Soit $\prod$ le produit cartésien. Chaque image I^n $(n \in [\![1, \mathcal{N}]\!])$ multivaluée (avec plusieurs bandes) peut être vue comme une fonction :*

$$\begin{array}{rccc} I^n : & [\![1,\mathcal{W}]\!] \times [\![1,\mathcal{H}]\!] & \rightarrow & \mathbb{Z}^{\mathcal{B}} \\ & (x,y) & \mapsto & I^n(x,y) = \prod_{b=1}^{\mathcal{B}} I_b^n(x,y) \end{array} \tag{1}$$

Définition 5 *Soit $\mathcal{S}^v$ le jeu de données construit à partir de la série d'images. $\mathcal{S}^v$ est un ensemble de séquences défini comme :*

$$\mathcal{S}^v = \{\!\{< I^1(x,y), \cdots, I^{\mathcal{N}}(x,y) > | x \in [\![1,\mathcal{W}]\!], y \in [\![1,\mathcal{H}]\!]\}\!\} \tag{2}$$

Ainsi, chaque séquence est une série de t-uplets (PIR,R,V,NDVI).

Finalement, il est nécessaire de discrétiser les valeurs des bandes afin de réduire le nombre d'items possibles pour l'étape d'extraction de motifs. Pour cela, $\mathcal{B}$ jeux de données $D_{b\ (1 \leqslant b \leqslant \mathcal{B})}$ sont tout d'abord créés. Ces jeux de données contiennent toutes les valeurs apparaissant dans la série sur la bande b correspondante. Formellement, chaque jeu de données est défini de la façon suivante :

$$D_b = \forall (x,y) \in [\![1,\mathcal{W}]\!] \times [\![1,\mathcal{H}]\!] : \biguplus_{n=1}^{\mathcal{N}} I_b^n(x,y) \tag{3}$$

Puis, chaque jeu de données D_b est discrétisé grâce à l'algorithme K-MEANS (MacQueen (1967)) afin d'obtenir K groupes de valeurs. Pour une question de lisibilité, ces groupes sont ordonnés et nommés en fonction de la valeur de leurs centroides, c'est-à-dire en fonction de la valeur moyenne du groupe. Ainsi, le cluster PIR_1 correspond à la première tranche de PIR, *i.e.,* le groupe de PIR contenant les plus faibles valeurs, et PIR_K correspond à la dernière tranche de PIR, *i.e.,* le groupe de PIR contenant les valeurs les plus élevées. La définition 6 détaille la création du jeu de données $\mathcal{S}$ utilisé pour l'extraction de motifs séquentiels.

Définition 6 *Soit $Clus$ la fonction associant les $\mathcal{B}$ valeurs décrivant un pixel (*i.e.*, un quadruplet $(pir, r, v, ndvi)$) aux $\mathcal{B}$ tranches correspondantes calculées par l'algorithme* K-MEANS. *Le jeu de données $\mathcal{S}$ est défini par :*

$$\mathcal{S} = \{\!\{< Clus(I^1(x,y)), \cdots, Clus(I^{\mathcal{N}}(x,y)) > | x \in [\![1,\mathcal{W}]\!], y \in [\![1,\mathcal{H}]\!]\}\!\} \tag{4}$$

Pour chaque pixel, nous disposons donc d'une séquence de valeurs discrètes, comme par exemple $(PIR_1; R_6; V_3; NDVI_{16}) \rightarrow \cdots \rightarrow (PIR_{12}; R_3; V_{14}; NDVI_{19})$ où $(PIR_1; R_6; V_3; NDVI_{16})$ signifie que les valeurs de ce pixel dans la première image font partie du premier groupe de proche infra-rouge, dans le sixième groupe de rouge, dans le troisième groupe de vert et dans le 16ème groupe de $NDVI$.

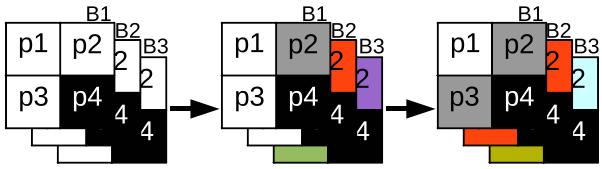


FIG. 3 – *Une série de 3 images avec 4 pixels, décrits sur 3 bandes.*

5 Extraction et visualisation des motifs issus des SITS

Le prétraitement décrit en section 4 produit des données où chaque pixel est décrit sur plusieurs mesures. Considérons les trois images de quatre pixels données en Figure 3. Nous voulons en extraire les motifs d'évolution. Cela signifie qu'un ensemble de pixels assez significatif doit avoir le même "comportement" pour qu'un motif en soit déduit. Dans notre exemple, chaque pixel est décrit sur 3 valeurs (les bandes $B1$ à $B3$). Avec un support minimum de 100 %, il n'y a pas de motif fréquent. Avec un support de 50 % on trouve deux motifs :

1. $< (B1, blanc; B2, blanc)\ (B1, gris; B2, rouge) >$. Ce motif correspond au pixel $p2$ sur les images 1 et 2 (ou 3) et $p3$ sur les images 1 (ou 2) et 3.
2. $< (B1, blanc; B2, blanc)\ (B1, blanc; B2, blanc) >$ ($p1$ et $p3$, images 1 and 2).

Notons que les motifs peuvent être fréquents même si les images qui les supportent ne sont pas les mêmes d'un pixel à l'autre (comme l'indique la définition 3).

5.1 Extraction de motifs séquentiels dans les séries d'images

La principale difficulté liée à la découverte d'évolutions fréquentes à l'aide de motifs séquentiels dans les SITS vient du fait que la très grande majorité des motifs extraits sont "plats", c'est-à-dire qu'ils n'expriment aucune évolution. En d'autres termes, ces motifs fréquents contiennent le même item, répété plusieurs fois. Par exemple, le motif $< (arbre)(arbre)(arbre) >$ qui traduit le fait que la plupart des pixels contiennent une valeur correspondant à des arbres sur trois images de la série. Ces motifs ne sont pas très informatifs. Dans les SITS, les motifs avec les supports les plus élevés correspondent très souvent à des zones géographiques qui n'évoluent pas. Pourtant, les pixels qui se "comportent" de manières similaires peuvent aider à découvrir des évolutions significatives dans les SITS. *Le défi consiste donc à extraire des motifs qui respectent un support minimum tout en exprimant une évolution*.

Il y a plusieurs approches, naïves ou élaborées, pour répondre à ce problème mais elles ont toutes leurs inconvénients :

– Un première approche consisterait à supprimer les pixels qui n'évoluent pas. Ce n'est pas suffisant car les pixels restant peuvent encore supporter des motifs plats (par exemple en gardant la même valeur sur toutes les images, sauf sur la dernière).
– On peut également baisser le support minimum jusqu'à éliminer les motifs les plus fréquents (qui sont plats) et trouver des motifs d'évolution. Le fait de baisser le support est une difficulté bien connue en extraction de connaissances car elle s'accompagne souvent

d'une explosion combinatoire qu'il est préférable d'éviter. De plus, le nombre de motifs extraits serait très difficile à gérer.
- Dans (*anonymisé*) nous avons proposé une solution qui interdit, dans le processus d'extraction des motifs séquentiels, les motifs qui présentent deux valeurs successives identiques. Cette approche fonctionne mais elle est limitée car un motif plat peut prendre ses valeurs sur plusieurs bandes. Par exemple, le motif $< (arbre, b1)\ (arbre, b2) >$ correspond à la valeur des arbres sur la bande 1 puis sur la bande 2 (étant donné que $(arbre, b1) \neq (arbre, b2)$, ce motif plat est accepté).
- Finalement, une solution naïve et extrême, consisterait à supprimer les items très fréquents du processus d'extraction. Dans ce cas, certains motifs importants seraient ignorés. Si l'item $(arbre, b1)$ est très fréquent mais qu'il apparait dans le motif $< (arbre, b1)\ (urbain, b3) >$ alors il ne faut pas l'exclure du processus d'extraction.

Compte tenu des limites de ces approches, nous proposons un principe d'extraction des motifs d'évolution qui est basé sur ALGOMS[4]. Ce principe consiste à éviter, dans le processus d'extraction des motifs séquentiels, la succession de deux items très fréquents. Cela implique l'utilisation d'un support maximum (qui s'ajoute au support minimum). Au delà de sa capacité à éviter les motifs plats, ce principe réduit considérablement l'espace de recherche.

Définition 7 *Soit $maxSup$ un support de fréquence maximum et H l'ensemble des items dont le support est supérieur à $maxSup$ (i.e. $H = \{i \in \mathcal{I} | support(i) > maxSupp\}$). Soit $minSup$ le support de fréquence minimum et FI l'ensemble des items dont le support est supérieur à $minSup$. Le* facteur de reduction h *qui s'applique à FI en fonction de $maxSup$ est donné par $h = |FI|/|FI \backslash H|$.*

Un nombre d'items réduit permet de diminuer l'espace des combinaisons possibles. Le théorème 1 fournit une borne sur cette réduction. Pour simplifier l'écriture, nous considérons que l'itemset vide fait partie des combinaisons possibles. De plus, ce raisonnement est limité aux séquences de longueur paire, mais il peut-être généralisé à tous les types de séquences.

Théorème 1 *Soit n le nombre d'items fréquents et h le facteur de réduction introduit dans la définition 7. Soit S_t l'ensemble des séquences de longueur t possibles et $P_{[0..t]} = \bigcup_{j=0}^{t} S_j$ l'ensemble des séquences possibles de longueur 0 à t. Soit S'_t l'ensemble des séquences possibles de longueur t qui ne présentent pas d'occurrences successives d'items avec un support supérieur à $maxSup$. Soit $P'_{[0..t]} = \bigcup_{j=0}^{t} S'_j$ l'ensemble des séquences possibles de ce type et de longueur comprise entre 0 et t. Alors $|P_{[0..t]}| = \sum_{j=0}^{t} 2^{nj}$ et $|P'_{[0..t]}| = \sum_{j=0}^{t} \frac{2^{nj}}{h^{\frac{j}{2}}}$ (et $|P'_{[0..t]}| < |P_{[0..t]}|$).*

Preuve. Commençons par un rappel sur les propriétés des motifs séquentiels. $|PI|$, le nombre de motifs séquentiels possibles avec n items fréquents est donné par $|PI| = 2^n - 1$ (ou $|PI| = \sum_{k=1}^{n} C_n^k = 2^n$ si on ne considère pas l'itemset vide). $|S_t|$, le nombre de séquences possibles de longueur t est donnée par $|S_t| = 2^{nt}$. En effet, le nombre de combinaisons possibles pour S_t est donné par : $\underbrace{2^n \times 2^n \times \ldots \times 2^n}_{t\ fois}$.

[4] Algorithme anonymisé pour la soumission. Toutefois, ce principe peut s'appliquer à n'importe quel algorithme d'extraction de motifs séquentiels

Et $|P_{[0..t]}|$, le nombre de motifs séquentiels possibles de longueur 0 à t est donné par : $|P_{[0..t]}| = |\bigcup_{j=0}^{t} S_t| = \sum_0^t 2^{nj}$

Donnons maintenant une borne supérieure au nombre de motifs séquentiels quand deux items de H ne peuvent pas être consécutifs. $|PI \backslash H|$, le nombre d'itemsets possibles composés d'items i tels que $minSupp < support(i) < maxSup$ est donné par $|PI \backslash H| = \frac{|PI|}{h} = \frac{2^n}{h}$.

Le nombre de combinaisons possibles pour S_t, les séquences de longueur t sans occurrences contigües d'items de H est donné par : $\underbrace{2^n \times \frac{2^n}{h} \times 2^n \times \frac{2^n}{h} \times \ldots \times 2^n \times \frac{2^n}{h}}_{\frac{t}{2}\ fois}$

Donc $|S'_t| = (2^n \times \frac{2^n}{h})^{\frac{t}{2}} = \frac{2^{2n\frac{t}{2}}}{h^{\frac{t}{2}}} = \frac{2^{nt}}{h^{\frac{t}{2}}}$ et le nombre de séquences possibles de ce type avec une longueur entre 0 et t est donné par : $|P'_{[0..t]}| = |\bigcup_{j=0}^{t} S'_j| = \sum_0^t \frac{2^{nj}}{h^{\frac{j}{2}}}$ □

Si l'impact de h sur le nombre d'itemsets possibles est significatif, il devient crucial pour les motifs séquentiels. Nos expérimentations montrent que ce filtre permet d'accélérer les temps de calcul d'au moins un ordre de grandeur.

5.2 Visualiser un ensemble de motifs séquentiels extraits

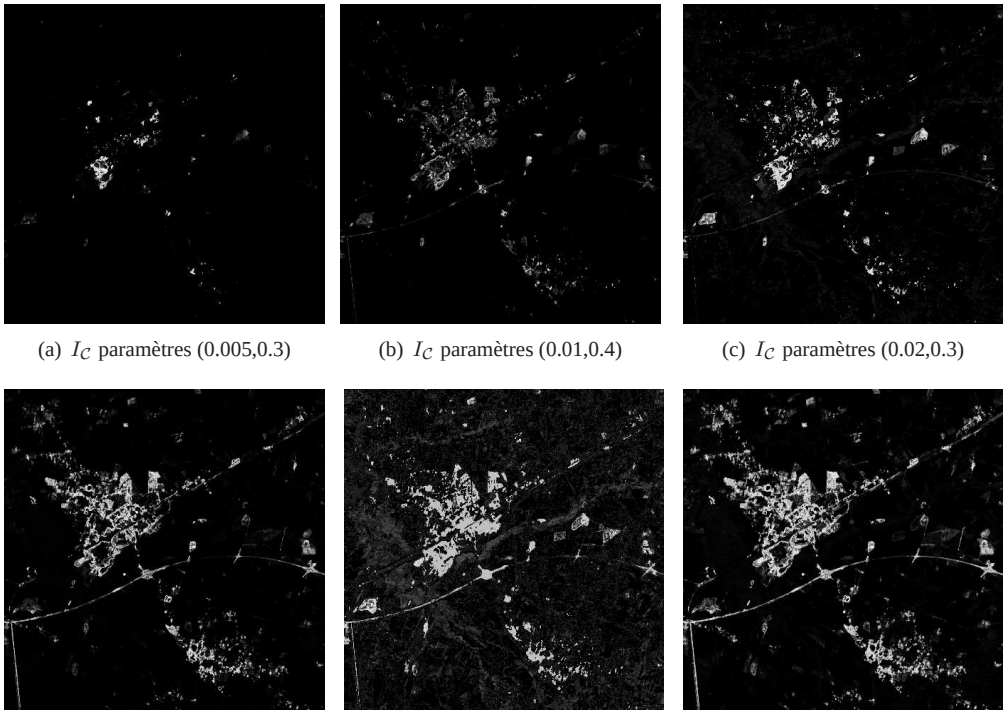
Le but de cette section est de proposer une méthode permettant de visualiser les motifs extraits de la SITS. Le grand nombre de motifs extraits par les algorithmes de fouille de données usuels est un problème important, qui est d'ailleurs traité dans de nombreuses contributions (Pei et al., 2004; Soulet et Crémilleux, 2008; Jeudy et Boulicaut, 2002). Dans le cas de l'extraction de motifs à partir d'images satellites, nous disposons d'un avantage pour la visualisation de ces motifs, de par le fait que les séquences représentent l'évolution d'une zone géographique. L'idée de notre méthode de visualisation est d'utiliser la position spatiale des séquences afin de construire une carte représentant les motifs extraits. Nous pouvons ainsi construire une image $\mathcal{H} \times \mathcal{W}$, sur laquelle chaque séquence participant au support d'un ou plusieurs motifs fréquents peut être affichée. De plus, en assignant des niveaux de gris progressifs en fonction de la participation d'une séquence aux différents motifs, il est possible de visualiser sur une seule carte, un ensemble de motifs extraits. A notre connaissance, *il s'agit de la première méthode permettant de donner une représentation synthétique d'un ensemble de motifs séquentiels extraits d'une série d'images.* La Définition 8 présente la "fréquence de contribution" d'une séquence à un ensemble de motifs séquentiels ; plus il y a de motifs supportés par une séquence, plus sa fréquence de contribution est élevée.

Définition 8 *Soit $F^{\mathcal{S}}$ l'ensemble des motifs séquentiels fréquents dans $\mathcal{S}$. La fréquence de contribution d'une séquence $s_{x,y} =< Clus(I^1(x,y)), \cdots, Clus(I^n(x,y)) >$ dans $F^{\mathcal{S}}$ est définie par :*

$$\mathcal{C}(s_{x,y}) \quad = \quad \frac{|\{s_F \prec s_{x,y} \mid s_F \in F^{\mathcal{S}}\}|}{|F^{\mathcal{S}}|} \tag{5}$$

Définition 9 *Soit $\mathcal{C}(s_{x,y})$ la fréquence de contribution d'une séquence $s_{x,y}$. L'image de fréquence de contribution $I_{\mathcal{C}}$ est définie par :*

$$\begin{array}{rccc} I_{\mathcal{C}} : & [1,\mathcal{W}] \times [1,\mathcal{H}] \subset \mathbb{N}^2 & \longrightarrow & [0,1] \subset \mathbb{R} \\ & (x,y) & \longmapsto & I_{\mathcal{C}}(x,y) = \mathcal{C}(s_{x,y}) \end{array} \tag{6}$$



(a) $I_{\mathcal{C}}$ paramètres (0.005,0.3) (b) $I_{\mathcal{C}}$ paramètres (0.01,0.4) (c) $I_{\mathcal{C}}$ paramètres (0.02,0.3)

(d) $I_{\mathcal{C}}$ paramètres (0.05,0.5) (e) $I_{\mathcal{C}}$ paramètres (0.07,0.3) (f) $I_{\mathcal{C}}$ paramètres (0.07,0.5)

FIG. 4 – *Images $I_{\mathcal{C}}$ pour différents paramètres d'extraction (support minimum, support maximum). Les pixels blancs correspondent aux valeurs élevées de $\mathcal{C}(s_{x,y})$.*

Étant donnée la définition 8, $\mathcal{C}(s_{x,y})$ peut être calculée pour chaque séquence $s_{x,y}$ (*i.e.*, pour chaque coordonnée (x, y)). La définition 9 détaille la formation de l'image $I_{\mathcal{C}}$ représentant la contribution de chaque séquence à l'ensemble de motifs séquentiels. Une fois cette représentation synthétique obtenue, l'expert dispose d'une carte lui permettant d'identifier rapidement les différentes zones impliquées dans l'ensemble de motifs extraits. Par exemple, si l'expert identifie principalement des zones agricoles, il peut alors parcourir l'ensemble des motifs extraits à la recherche de rotations de cultures.

6 Expérimentations

Nous disposons de 35 images d'une résolution de 202 500 pixels (450x450). Elles sont extraites de la base de données Littoral du programme KALIDEOS et se situent dans le sud-ouest de la France. Nous avons extrait de cette base de données une série de 35 images SPOT-1, SPOT-2 et SPOT-4 (illustrée en figure 2). Une fois ces images prétraitées (C.f. la section 4), chaque pixel est décrit par quatre attributs (PIR,R,V,NDVI) et les séries correspondantes contiennent 28 millions de valeurs.

6.1 Visualisation d'ensembles de motifs extraits des SITS

La Figure 4 présente différents exemples d'images $I_{\mathcal{C}}$ calculées pour différents $F^{\mathcal{S}}$. De façon non surprenante, le support minimum influe de façon directe sur le nombre de séquences supportant les différents motifs. Aussi, lorsque ce support est faible, les motifs extraits sont supportés par peu de séquences et inversement. Au delà de cette observation évidente, la Figure 4(a) montre que les séquences (et donc les pixels des images $I_{\mathcal{C}}$) supportant les motifs avec de faibles valeurs de supports correspondent principalement à des zones industrialisées. La Figure 4(c), quant à elle, contient un peu plus de séquences supportant l'ensemble des motifs extraits et correspond à des zones urbaines ou asséchées. Enfin, la Figure 4(e) laisse apparaître, en plus des zones urbaines, des zones humides (marais) en gris foncé. Cette image nous informe également sur un autre point : la constance du niveau de gris sur les zones urbaines indique que ces régions supportent souvent en même temps les différents motifs de l'ensemble.

6.2 Expressivité (évolution) des motifs extraits

Les images construites selon le principe décrit en 5.2 fournissent aux experts une visualisation intuitive des types de motifs extraits en fonction des supports minimum et maximum. De ces ensembles, trois motifs ont été extraits par un expert géographe ; nous détaillons ci-dessous leur interprétation géographique :

1. Le motif <**(IR,1) (NDVI,20)**>, qui correspond à des zones marécageuses. Le niveau oscille entre un niveau très faible de proche infra-rouge quand les zones sont recouvertes d'eau en hiver et un fort niveau NDVI en été, car la végétation apparait, et est de plus très irriguée et donc très chlorophyllienne.
2. Le motif <**(R,17) (R,18 ; NDVI,3)**> qui correspond à des zones urbaines s'étant densifiées (augmentation de 25% de la réflectance). La présence d'urbain est de plus corroborée par le niveau final très faible de NDVI indiquant une quasi-absence de végétation. L'intérêt d'une approche sur plusieurs bandes est donc particulièrement bien illustré par ce motif.
3. Le motif <**(NDVI,2) (G,20) (NDVI,1)**> correspond à des zones industrielles s'étant densifiées. Ici, la diminution de NDVI indique un recul du couvert végétal et le niveau maximum atteint dans le vert corrobore la présence de toits plats (type tôle ondulée) caractéristiques des zones industrielles.

6.3 Impact du support maximum sur les temps de calcul

La notion de support maximum introduite en 5.1 a une grande influence sur le nombre d'items fréquents qui ne sont pas autorisés à apparaitre de manière contigüe (et donc sur le nombre total de motifs évalués). La figure 5 (gauche) indique le nombre d'items qui sont combinés dans les motifs. Par exemple, avec un support maximum de 75%, le nombre d'items i dont le support est compris entre 50% et 75% est de 37. Cela correspond à un facteur de réduction $h = 1,7$ par rapport au nombre original. Les temps de réponses avec un support maximum de 80%, 75% et 70% sont reportés en figure 5 (droite) et sont comparés aux temps de réponses obtenus avec les items sans support maximum ("Original"). On peut y observer

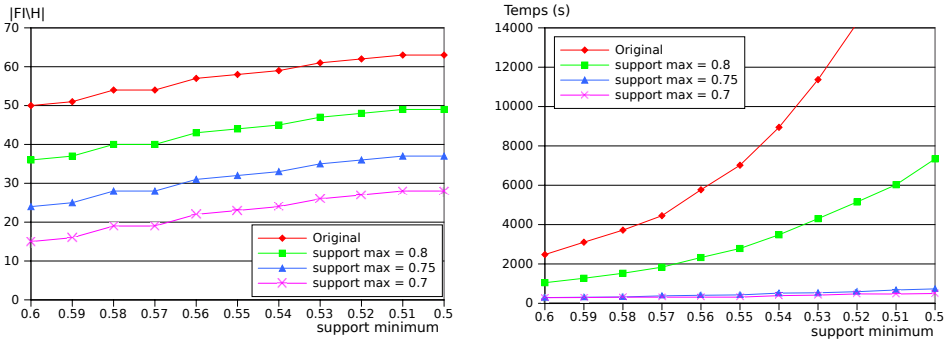


FIG. 5 – *Nombre d'items dans* $|FI\backslash H|$ *(gauche) et temps de réponses (droite).*

que l'influence de h est très importante sur les temps de réponses, comme attendu et décrit en 5.1. Le principe du support maximum est très adapté à la fouille de SITS, dans la mesure où les motifs plats sont évités et les performances dépassent celle d'une application directe d'environ un ordre de grandeur. Lors de nos expérimentations, nous avons pu constater qu'avec un support minimum de 50% et sans support maximum, l'extraction pouvait prendre plusieurs heures. Avec un support minimum de 10% l'extraction pouvait prendre plusieurs semaines. En revanche, avec un support minimum de 10% et un support maximum de 50%, l'extraction se termine en quelques minutes.

7 Conclusion

Les motifs séquentiels sont très adaptés à la découverte et la description d'évolutions dans les données. Toutefois, quand il s'agit d'extraire des motifs d'évolution dans les SITS, le défi consiste à filtrer le très grand nombre de motifs plats, afin de ne pas saturer les experts. Dans cet article, nous avons proposé i) une méthode permettant d'extraire les motifs d'évolution dans les SITS, ii) un principe permettant de filtrer les motifs plats et iii) une technique de visualisation qui permet de localiser les zones d'évolution de manière rapide et intuitive. Nos expérimentations montrent la pertinence de notre approche et des motifs extraits.

Références

Agrawal, R. et R. Srikant (1995). Mining Sequential Patterns. In *Proceedings of the 11th International Conference on Data Engineering (ICDE'95)*, pp. 3–14.

Bruzzone, L. et D. Prieto (2000). Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing 38*(3), 1171–1182.

Campbell, J. B. (2007). *Introduction to Remote Sensing*. The Guilford Press.

Coppin, P., I. Jonckheere, K. Nackaerts, B. Muys, et E. Lambin (2004). Digital change detection methods in ecosystem monitoring : a review. *International Journal of Remote Sensing 25*, 1565–1596.

Howarth, P., J. Piwowar, et A. Millward (2006). Time-Series Analysis of Medium-Resolution, Multisensor Satellite Data for Identifying Landscape Change. *Photogrammetric engineering and remote sensing 72*(6), 653–663.

Jensen, J. R. (2007). *Remote Sensing of the Environment : An Earth Resource Perspective*. Prentice Hall.

Jeudy, B. et J. Boulicaut (2002). Using condensed representations for interactive association rule mining. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 225–236.

Julea, A., N. Méger, C. Rigotti, M.-P. Doin, C. Lasserre, E. Trouvé, P. Bolon, et V. Lazarescu (2010). Extraction of Frequent Grouped Sequential Patterns from Satellite Image Time Series. In *Proceedings of the International Geoscience and Remote Sensing Symposium, 2010. IGARSS 2010. IEEE International. International Geoscience and Remote Sensing Symposium, 2010. IGARSS 2010.*, Honolulu USA, pp. 4.

Julea, A., N. Méger, et E. Trouvé (2006). Sequential patterns extraction in multitemporal satellite images. In *Workshop on Practical Data Mining (in conjunction with PKDD)*, pp. 96–99.

Julea, A., N. Méger, E. Trouvé, et P. Bolon (2008). On extracting evolutions from satellite image time series. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Volume 5, pp. 228–231.

Kennedy, R. E., W. B. Cohen, et T. A. Schroeder (2007). Trajectory-based change detection for automated characterization of forest disturbance dynamics. *Remote Sensing of Environment 110*(3), 370–386.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Pei, J., G. Dong, W. Zou, et J. Han (2004). Mining condensed frequent-pattern bases. *Knowledge and Information Systems (KAIS) 6*(5), 570–594.

Soulet, A. et B. Crémilleux (2008). Adequate condensed representations of patterns. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML / PKDD)*, pp. 20–21. Springer-Verlag.

Summary

Satellite Image Time Series (SITS) provide us with precious information on land cover evolution. By studying these series of images we can understand the changes of specific areas but also discover global phenomena that spread over larger areas. However, discovering these evolution patterns implies to consider two main challenges, related to the characteristics of SITS and the domain's constraints. We propose a SITS mining framework that allows for discovering these patterns despite these constraints and characteristics. Our proposal is inspired from sequential pattern mining and provides a relevant visualisation principle.