HAL
open science

Recherche de motifs et cartographie des surfaces agricoles. Des relevés terrain aux données satellitaires : application au Mali

Elodie Vintrou, Yoann Pitarch, Agnès Bégué, Maguelonne Teisseire

To cite this version:

Elodie Vintrou, Yoann Pitarch, Agnès Bégué, Maguelonne Teisseire. Recherche de motifs et cartographie des surfaces agricoles. Des relevés terrain aux données satellitaires : application au Mali. Revue Internationale de Géomatique, 2011, 21 (4), pp.469-488. 10.3166/rig.21.469-488 . hal-00682846

HAL Id: hal-00682846

https://hal.science/hal-00682846

Submitted on 27 Mar 2012

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

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

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

Recherche de motifs et cartographie des surfaces agricoles

Des relevés terrain aux données satellitaires : application au Mali

Elodie Vintrou * — **Yoann Pitarch**** — **Agnès Bégué*** — **Maguelonne Teisseire**,***

* *CIRAD, UMR TETIS F-34093 Montpellier, France*

** *LIRMM-CNRS-UM2 F-34095 Montpellier, France*

*** *Cemagref, UMR TETIS F-34093 Montpellier, France*

{nom}@teledetection.fr, pitarch@lirmm.fr

RÉSUMÉ. *La cartographie automatique de territoires ruraux est un outil essentiel dans le contexte sociétal actuel (e.g., analyse des risques de famine, étude des risques liés à la déforestation). Dans cet article, nous présentons une approche préliminaire de caractérisation des paysages ruraux et de leurs systèmes de culture à partir de techniques de fouille de données (recherche d'itemsets fréquents). Cette méthode permet de coupler des données de relevés terrain aux indicateurs extraits des images satellitaires. Cette approche a été mise en œuvre sur des données associées au Mali en collaboration avec des experts du domaine posant les premières bases d'une méthode originale d'extraction de motifs à partir de données complexes.*

ABSTRACT. *Countryside automatic cartography is a real and decisive challenge in the current societal context (e.g., the famine risk analysis, the deforestation consequence analysis). In this paper, we propose a preliminary approach allowing the landscape characterization. More precisely, an itemset-based-technique is developed to extract crop types features. One of the main strengths of the proposed methodology is to combine both indicators extracted from satellite image and data collected from a site survey. The approach was run on data associated to Mali in collaboration with some domain experts.*

MOTS-CLÉS : *Fouille de données, itemsets fréquents, image satellite, occupation du sol*

KEYWORDS: *Data Mining, Frequent Itemsets, Satellite Image, Land Cover*

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

1. Introduction

Motivés par des problèmes d'Aide à la Décision, les chercheurs de différentes communautés (Intelligence Artificielle, Statistiques, Bases de Données ...) se sont intéressés à la conception et au développement d'une nouvelle génération d'outils permettant d'extraire automatiquement de la connaissance de grandes bases de données. Ces outils, techniques et approches sont le sujet d'un thème de recherche connu sous le nom de *Knowledge Discovery in Databases* ou KDD (Extraction de Connaissances dans les Bases de Données) dont le Data Mining (Fouille de Données) est une étape. Elles sont utilisées dans de nombreux domaines d'applications. Les exemples les plus courants sont les compagnies d'assurance, les compagnies bancaires (crédit, prédiction du marché, détection de fraudes), le marketing (comportement des consommateurs, mailing personnalisé), la recherche médicale (aide au diagnostic, au traitement, surveillance de population sensible), les réseaux de communication (détection de situations alarmantes, prédiction d'incidents), l'analyse de données spatiales, etc.

La fouille de données peut être définie par « *Processus non trivial permettant l'extraction automatique de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles et compréhensibles* » (Fayyad *et al.*, 1996). Bien que le terme de fouille de données représente la découverte de connaissances, il ne constitue en fait qu'une seule des étapes du KDD, qui comprend globalement trois étapes : la préparation des données, l'extraction des données (Data Mining) et leur interprétation.

La première étape consiste à sélectionner uniquement les données potentiellement utiles de la base (opération de filtrage), sur lesquelles on effectue une phase de pré-traitement (gestion des données manquantes ou invalides). Ensuite, les données obtenues passent par une phase de formatage, afin de les préparer au processus de Data Mining. Finalement, la dernière étape est une étape d'analyse et d'interprétation de la connaissance extraite par la fouille de données, pour la rendre lisible et compréhensible par l'utilisateur. Les besoins variés nécessitent des approches différentes telles que la classification, la recherche de corrélations, la segmentation ou encore la détection de déviations.

Les travaux réunissant les chercheurs des communautés télédétection et fouille de données sont récents et correspondent essentiellement à des approches utilisant les arbres de décision (dos Santos Silva *et al.*, 2005; Aksoy *et al.*, 2004) ou la détection de changement au sein de séries temporelles (Romani *et al.*, 2010; dos Santos Silva *et al.*, 2005). Il existe également des propositions de recherche de motifs comme outil d'exploration des données satellitaires. Nous pouvons citer (Julea *et al.*, 2011; Petitjean *et al.*, 2010) qui extraient des patrons de comportement ou identifient des profils de changement.

Notre objectif est d'utiliser des techniques de fouille de données permettant de coupler des données hétérogènes pour faire une cartographie de l'occupation des sols

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

à partir de relevés terrain, de données environnementales et d'informations issues d'images satellitaires. Ce projet est donc une alternative pertinente pour définir un mécanisme d'apprentissage basé sur des données multi-sources (données spectrales, texturales et temporelles), des données environnementales (climat, relief, type de sol,. . .) et des données de terrain, et mettre en évidence des relations qui n'auraient pas pu être identifiées autrement. Pour cela, nous proposons d'adapter un mécanisme de recherche de motifs séquentiels multidimensionnels, comme proposé dans (Plantevit *et al.*, 2010), à la fouille de séries d'images satellitaires et des données pouvant être mises en relation (informations externes).

Un des enjeux de l'application de ces techniques en télédétection réside dans le fait que les données relatives à une même réalité terrain sont par nature multi-sources, puisqu'elles résultent du croisement de données issues de différents capteurs, de relevés terrain et bases de données externes (Leenhardt *et al.*, 2005). Leur description se fait par ailleurs dans plusieurs dimensions et combine des informations spectrales, spatiales et temporelles. Pour cette raison, nous nous tournons vers des algorithmes d'extraction d'itemsets multidimensionnels (Pinto *et al.*, 2001), qui peuvent prendre en compte le caractère multidimensionnel des données.

Le jeu de données utilisé dans cette étude a été acquis dans la région soudano-sahélienne du Mali, qui présente l'intérêt d'être une zone particulièrement complexe pour laquelle la cartographie de l'occupation du sol par télédétection donne des résultats mitigés (Hansen *et al.*, 2000; Fritz *et al.*, 2010). Les principales raisons évoquées pour expliquer cette difficulté est que la zone soudano-sahélienne est une zone de transition éco-climatique présentant une grande variabilité spatiale des systèmes agricoles et des calendriers culturaux. De plus, le parcellaire agricole y est généralement petit, l'hétérogénéité inter et intra-parcellaire est grande et les couverts végétaux, naturels et cultivés, ont le même cycle de croissance en raison du régime des pluies (Vintrou *et al.*, 2011). Nous proposons de tester deux jeux de données d'images satellitaires : un jeu de données monodate à haute résolution spatiale (2.5 m) et un jeu de données mensuelles à faible résolution spatiale. Ces deux jeux de données sont représentatifs des jeux de données actuellement disponibles en Observation de la Terre.

Les résultats présentés dans cet article mettent en œuvre un jeu de données restreint d'imagerie satellitaire et de données exogènes. Cette approche préliminaire, centrée sur l'apprentissage du système, nous a semblée indispensable pour assurer des bases communes aux deux communautés engagées dans ce travail, les agronomes et les spécialistes de fouille de données, et ainsi initier des collaborations de recherche plus pointues.

Dans les sections suivantes, nous décrivons tout d'abord les données étudiées (section 2) pour ensuite présenter les définitions associées aux motifs multidimensionnels

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

(section 3). Puis, nous décrivons en détail le processus d'extraction mis en œuvre (section 4) ainsi que l'analyse des premiers résultats obtenus (section 5).

2. Description des données

2.1. *Contexte et zone d'étude*

Le Mali est un pays d'Afrique de l'Ouest, autour de la latitude 14°N. Ce pays possède un gradient climatique Sud-Nord, qui varie de régions subtropicales à semi-arides, et s'étend plus au Nord vers des zones arides et désertiques.

Le Mali peut être considéré comme représentatif de la zone soudano-sahélienne, où la forte dépendance à l'agriculture pluviale entraîne une vulnérabilité aux changements climatiques et anthropiques. Par conséquent, afin de mieux connaître le phénomène de mousson en Afrique de l'Ouest et sa variabilité, et d'améliorer les prévisions des impacts de cette variabilité sur l'agriculture et la sécurité alimentaire, une des premières étapes nécessaires est une estimation fiable du domaine cultivé. Une attention particulière a été portée sur 3 zones, le long du gradient climatique malien (tableau 1).

2.2. *Les données terrain*

Des missions de terrain ont été effectuées au Mali de mai à novembre 2009, dans le but de caractériser les paysages agricoles soudano-sahéliens. Au total, 744 points GPS ont été enregistrés, et des paysans de chacune des régions étudiées ont été interrogés. Les relevés ont été effectués par transects, du centre du village vers l'extérieur. Chaque point relevé a été transformé en un polygone dont il est le centre et auquel a été affecté un type d'occupation du sol ("mil", "sorgho", "maïs", "coton", "végétation naturelle" principalement), comme indiqué figure 7.

2.3. *Les villages*

La base de données village a été fournie par l'Institut d'Economie Rurale de Bamako. Elle contient le nom des villages de l'ensemble du Mali, ainsi que leurs coordonnés géographiques comme illustré figure 1.

2.4. *Les données images*

2.4.1. *Les images SPOT*

Les images à haute résolution spatiale utilisées dans cette étude sont des images SPOT5 multispectrales (bandes dans le "vert", "rouge" et "proche infrarouge") à 2.5 m

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

Site détude (éco-climat)	Précipitations moyennes annuelles	Culture principale	Végétation naturelle majoritaire	surface culti-vée	Date SPOT
Cinzana (soudano-sahélien)	600 mm	mil et sorgho	végétation dégradée et sol nu	43%	14 nov. 2007
Koutiala (soudano-sahélien)	750 mm	mil, sorgho et coton	végétation ouverte et fermée	52%	20 nov. 2007
Sikasso (soudanien)	1 000 mm	mais, coton et fruits	végétation dense	40%	20 nov. 2007

Tableau 1. *Principales caractéristiques des trois zones d'étude et dates d'acquisition des images SPOT*

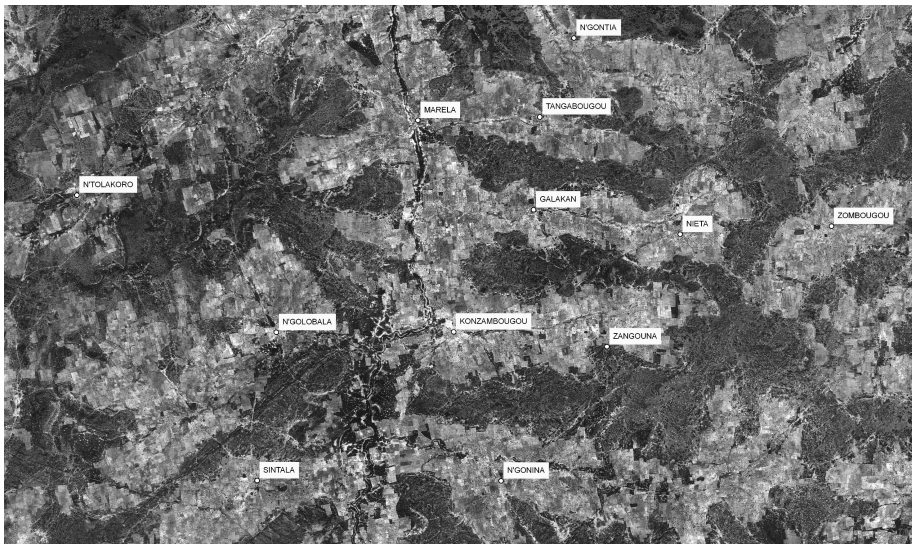

Figure 1. *Exemples de villages - zoom sur SPOT CINZANA (21/11/2007)*

de résolution, livrées orthorectifiées. Les images d'archive les plus récentes sur nos sites d'étude ont été acquises en novembre 2007 (tableau 1), ce qui correspond à la fin de la saison des pluies. Le décalage entre l'année d'acquisition des images (2007) et les relevés de terrain (2009) n'est pas un problème car l'occupation du sol a peu évolué en deux ans.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

2.4.2. *La série temporelle d'images MODIS*

Les images MODIS sont des images multispectrales à basse résolution spatiale (250 m), mais à large champ (couverture du Sud Mali en une scène) et à haute répétitivité (acquisition journalière). Dans cette étude nous avons utilisé les produits MOD13Q1/ V05 du LP-DAAC (NASA Land Process Distributed Active Archive Center) qui correspondent à des synthèses sur 16 jours d'indice de végétation (NDVI). Nous avons récupéré la série complète (24 images) de 2007 pour être en cohérence avec les dates d'acquisition des images SPOT. Sur ces 24 images, nous avons conservé pour chaque mois la meilleure image acquise (avec l'ennuagement le plus faible).

2.5. *Les données dérivées des images*

À partir de ces images sont calculés un certain nombre de descripteurs spectraux (indice de végétation) et spatiaux (indices de texture) des états de surface (occupation du sol) :

– *L'indice de végétation NDVI*. Les indices de végétation sont des combinaisons, linéaires ou non, de réflectances dans les bandes spectrales R (rouge), PIR (proche infrarouge) et MIR (moyen infrarouge). Ils permettent de caractériser le couvert végétal en terme de vigueur de la végétation. Parmi ces indices, le « *Normalized Difference Vegetation Index* » (Rouse, 1974) est de loin le plus utilisé. Il permet de caractériser l'activité photosynthétique de la surface, et donc de discriminer facilement des surfaces végétalisées (0,9 pour de la végétation verte et dense) des surface de sols nus (environ 0,1 pour sol nu). Ces propriétés font que le NDVI est fréquemment utilisé comme une mesure indirecte de la biomasse. La formule du NDVI est la suivante :

$$NDVI = \frac{PIR - R}{PIR + R}$$

– *Les indices de texture*. La variabilité spatiale d'une image est représentée par le concept de texture. Haralick élargit dans (Haralick, 1979) la définition en décrivant une texture comme un phénomène à deux dimensions : la première concernant la description d'éléments de base ou primitives (le motif) à partir desquels est formée la texture ; la deuxième dimension est relative à la description de l'organisation spatiale de ces primitives.

La matrice de co-occurrences (ou matrice de dépendance spatiale) est une des approches les plus connues et les plus utilisées pour extraire des caractéristiques de textures. Elle effectue une analyse statistique de second ordre de la texture, par l'étude des relations spatiales des couples de pixels (Haralick *et al.*, 1973; Haralick, 1979). Quatorze indices (définis par Haralick) qui correspondent à des caractères descriptifs des textures peuvent être calculés à partir de cette matrice.

Dans cette étude, chaque indice d'homogénéité, de variance, de dissimilarité et de contraste a été calculé sur une fenêtre de 15×15 pixels.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

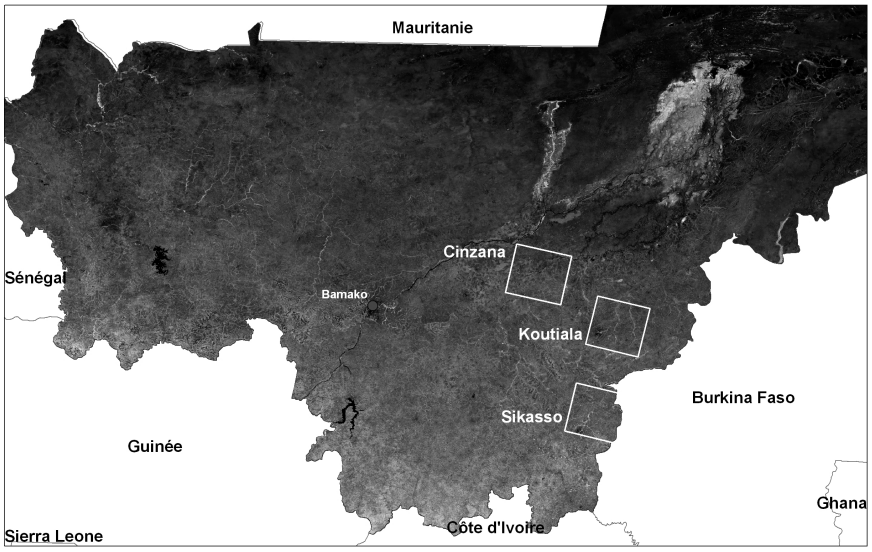

Figure 2. *Carte du Mali sur fond de série temporelle MODIS 2007 et emprise géographique des trois images SPOT de validation (Cinzana, Koutiala, Sikasso).*

3. Extraction d'itemsets séquentiels multidimensionnels

Le problème de la recherche de motifs séquentiels a été introduit par R. Agrawal dans (Agrawal *et al.*, 1995) et appliqué avec succès dans de nombreux domaines comme la biologie (Wang *et al.*, 2004; Salle *et al.*, 2009), la fouille d'usage du Web (Pei *et al.*, 2000; Masseglia *et al.*, 2008), la détection d'anomalie (Rabatel *et al.*, 2010), la fouille de flux de données (Marascu *et al.*, 2006) ou la description des comportements au sein d'un groupe (Perera *et al.*, 2009).

Des approches plus récentes (Julea. *et al.*, 2008) utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein des séries d'images satellites. Néanmoins, à notre connaissance, l'étude de la littérature ne fait état d'aucuns travaux sur l'application de techniques de recherche de motifs séquentiels couplant données externes et télédétection.

Dans cette section, nous introduisons les définitions relatives à la fouille d'itemsets séquentiels multidimensionnels et décrivons l'algorithme de fouille de données utilisé.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

3.1. *Itemsets et séquences multidimensionnels*

L'approche M^3SP (Plantevit *et al.*, 2010) permet l'extraction de motifs séquentiels multidimensionnels. Cette méthode est adaptée à notre contexte car elle permet d'analyser à la fois la dimension temporelle des séries d'images MODIS ainsi que les informations terrain associées. Les concepts proposés sont définis dans cette section.

Soit un ensemble $\mathcal{D}$ de dimensions et $\{\mathcal{D}_R, \mathcal{D}_A, \mathcal{D}_T, \mathcal{D}_I\}$ une partition de $\mathcal{D}$ dans laquelle $\mathcal{D}_R$ désigne les dimensions de référence, qui permettent de déterminer si une séquence est fréquente, $\mathcal{D}_A$ les dimensions d'analyse, sur lesquelles les corrélations sont extraites, et $\mathcal{D}_T$ les dimensions permettant d'introduire une relation d'ordre (généralement le temps). Les dimensions $\mathcal{D}_I$ sont les dimensions ignorées lors de la fouille.
Dans notre contexte, la dimension de référence ($\mathcal{D}_R$) peut être la dimension associée aux points terrains, les dimensions d'analyse ($\mathcal{D}_A$) peuvent inclure les dimensions décrivant les types d'occupation du sol, les indices de texture obtenus des pixels correspondant.

Pour chaque dimension $D_i \in D$, on note $Dom(D_i)$ son domaine de valeurs. À chaque domaine de valeurs $Dom(D_i)$ est associé une hiérarchie H_i, et l'on suppose que $Dom(D_i)$ contient une valeur particuliére notée $\top_i$ (la racine de la hiérarchie). Lorsqu'aucune hiérarchie de valeurs n'est définie sur une dimension D_i, nous considérons H_i comme un arbre de profondeur 1 dont la racine est $\top_i$ et dont les feuilles sont les éléments de $Dom(D_i) \setminus \{\top_i\}$. La figure 3 présente un exemple de hiérarchie de valeurs H_i pour la dimension site-étude et la figure 7 celle de la dimension type-utilisation-sol .

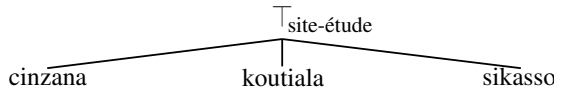

Figure 3. *La hiérarchie de valeurs H_i pour la dimension site-étude.*

Un *item multidimensionnel* $e = (d_1, d_2, \ldots, d_m)$ est un m-uplet défini sur les dimensions d'analyse $\mathcal{D}_A$, c'est- à-dire tel que $\forall i \in [1 \ldots m], d_i \in Dom(D_i)$ avec $D_i \in \mathcal{D}_A$ et $\exists d_i \in [1, \ldots, m]$ tel que $d_i \neq \top_i$. Par exemple, $e = (sorgho, cinzana)$ et $e' = (sorgho, \top_{site-etude})$ sont des items multidimensionnels qui décrivent des points terrain sur les dimensions d'analyse $\mathcal{D}_A = \{type - utlisation - sol, site - etude\}$. On définit une relation d'inclusion $\subseteq$ entre items multidimensionnels : un item multidimensionel $e = (d_1, d_2, \ldots, d_m)$ est inclus dans un item multidimensionnel $e' = (d'_1, d'_2, \ldots, d'_m)$ (noté $e \subseteq e'$) si $\forall i \in [1, \ldots, m]$, $d_i = d'_i$ ou est une spécialisation de d'_i dans H_i.

Dans l'exemple précédent, on a l'inclusion $(sorgho, cinzana) \subseteq (sorgho, \top_{site-etude})$ car $cinzana$ est une spécialisation de $\top_{site-etude}$ dans la hiérarchie $H_{site-etude}$.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

pt-id	date	type-utlisation-sol	site-étude	SPOT-NDVI-100
1	1	arachide	sikasso	très faible
2	1	mil	koutiala	faible
2	2	mil	koutiala	modéré
3	1	sorgho	koutiala	élevé

Tableau 2. *Base de données DB.*

pt-id	date	type-utlisation-sol	site-étude	SPOT-NDVI-100
2	1	mil	koutiala	faible
2	2	mil	koutiala	modéré

Tableau 3. *Bloc $B_{(mil,koutiala)}$.*

Un *itemset multidimensionnel* $i = e_1, e_2, \ldots, e_m$ est un ensemble non vide d'items multidimensionnels non deux à deux comparables par rapport à $\subseteq$ (c.- à-d., $\forall i, j \in [1, \ldots, m], e_i \nsubseteq e_j$ et $e_i \nsupseteq e_j$). On définit une relation d'inclusion $\subseteq$ entre itemsets multidimensionnels : un itemset i est inclus dans un itemset i' (noté $i \subseteq i'$) si pour chaque item a de i, il existe au moins un item a' de i' tel que $a \subseteq a'$.

Une *séquence multidimensionnelle* $s = \langle i_1, \ldots, i_n \rangle$ est une liste ordonnée non vide d'itemsets multidimensionnels.

Une relation de généralisation (ou spécialisation) entre séquences multidimensionnelles est définie. Une séquence $s = \langle i_1, i_2, \ldots, i_n \rangle$ est plus spécifique qu'une séquence $s' = \langle i'_1, i'_2, \ldots, i'_m \rangle$ s'il existe des entiers $1 \leq j_1 \leq \ldots \leq j_m \leq n$ tels que $s_{j_1} \subseteq s'_1, s_{j_2} \subseteq s'_2, \ldots, s_{j_m} \subseteq s'_m$.

Étant donnée une table relationnelle DB, on appelle *bloc* l'ensemble des n-uplets qui ont la même projection sur $\mathcal{D}_R$. Par exemple, le tableau 3 donne le bloc formé en ne gardant que les n-uplets de la table relationnelle DB donnée au tableau 2 dont la projection sur $\mathcal{D}_A = \{type - utilisation - sol, site - etude\}$ est $(mil, koutiala)$. Le *support* d'une séquence est le nombre de blocs qui contiennent cette séquence.

Étant donné un seuil σ_{min} de support minimum, le but de la recherche d'itemsets séquentiels multidimensionnels est de trouver toutes les séquences dont le support est supérieur ou égal à σ_{min}.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

4. Mise en œuvre

Un processus d'extraction de connaissances a été mis au point (figure 4). Il est constitué de quatre étapes : (1) chargement des données, (2) préparation des données, (3) fouille de données et (4) interprétation et validation des résultats.

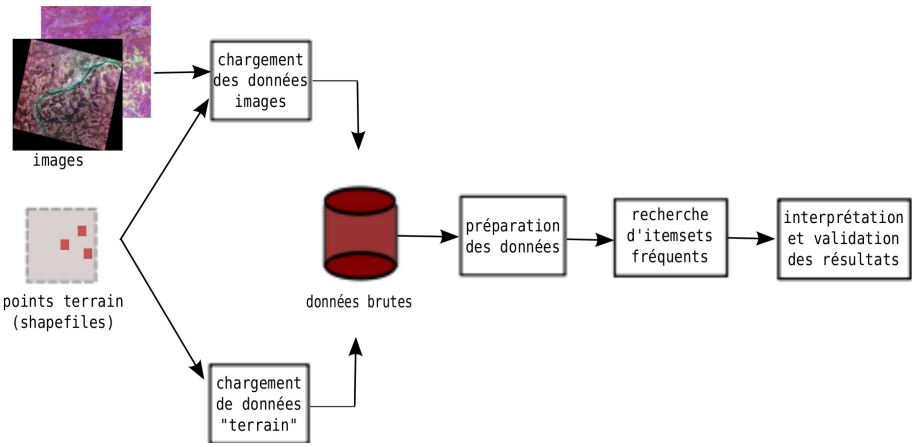

Figure 4. *Les différentes étapes du processus d'extraction de connaissances.*

Dans l'étape de chargement des données, les données sont collectées et placées dans une base de données (données "brutes"). Cinq indices images sont calculés à partir des images SPOT et MODIS : l'indice de végétation NDVI et les indices de texture variance, homogénéité, contraste et dissimilarité. A chaque point terrain est associée une valeur de chacun de ces indices, qui est calculée en prenant la moyenne de la valeur de l'indice sur les pixels contenus dans un polygone carré centré sur ce point (figure 5). Le choix de la taille du polygone pour les images SPOT a été régi par la taille des parcelles cultivées au Mali. Celle-ci est très variable suivant les régions, mais pour les points échantillonnés, un polygone de 100mx100m est toujours inclus dans une parcelle : le NDVI et la texture calculés au sein d'un polygone sont donc purs et concernent une seule occupation du sol. Pour les images MODIS, la taille de polygone choisie est de 1kmx1km, car à cette résolution, nous étudions non pas une parcelle cultivée mais un domaine cultivé à l'échelle d'un terroir.

L'étape de préparation des données a pour but de constituer l'ensemble d'apprentissage et de structurer les données afin de pouvoir être traitées par l'algorithme de fouille.

L'ensemble d'apprentissage est constitué en sélectionnant parmi l'ensemble des points terrain les points qui se situent sur une des images SPOT et qui correspondent à

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

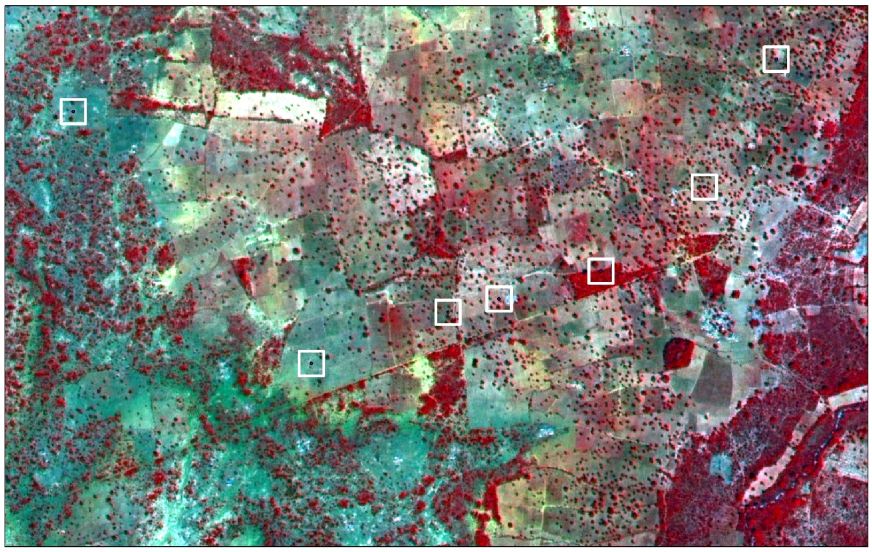

Figure 5. *Extraction des indices images pour chaque point terrain sur un polygone carré centré sur ce point (zoom sur une image SPOT du site de Koutiala, les carrés sont des polygones de 100 m×100 m centrés sur chacun des points terrain). © CNES 2009, Distribution SPOT Image*

Site d'étude	Culture	Non culture	Total
Cinzana	138	85	223
Koutiala	105	78	183
Sikasso	46	46	92

Tableau 4. *Les points terrain de l'ensemble d'apprentissage.*

des relevés en zone de culture ou non culture. On obtient 498 points dont la répartition par site d'étude et en culture / non culture est donnée par le tableau 4.

Chaque point de l'ensemble d'apprentissage est décrit par les valeurs qu'il prend dans un ensemble de dimensions D_i :

– id-pt : entier qui identifie chaque point terrain de manière unique

– date : estampille temporelle (constante pour les images SPOT et de 1 à 12 pour les images MODIS)

– site-étude : le nom du site d'étude

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

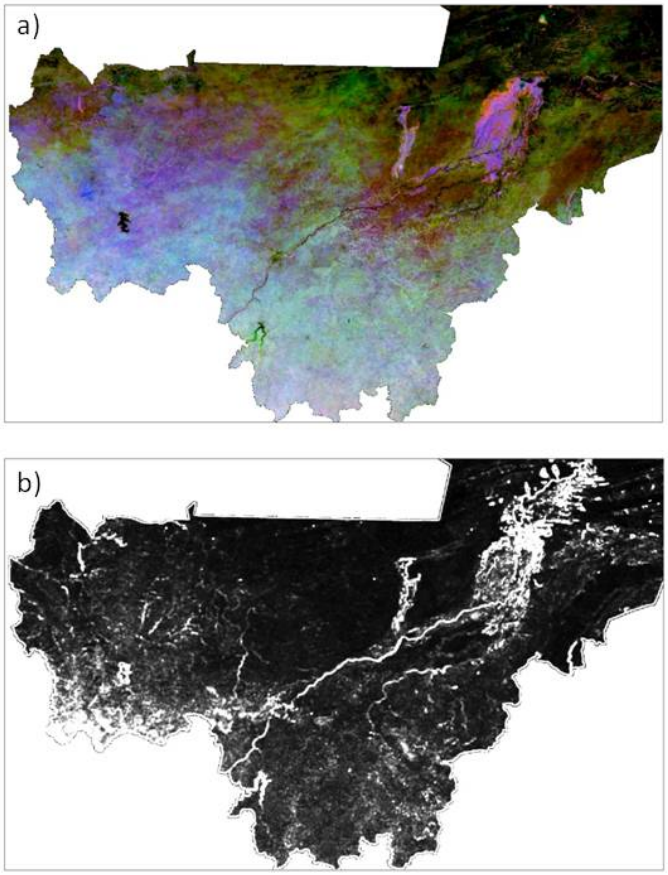

Figure 6. *Série temporelle 2007 a) MODIS NDVI, b) MODIS TEXTURE variance*

– type-occupation-sol : "culture" ou "non culture"

– nom-village : le nom du village le plus proche, pris parmi les relevés village

– distance-village : la distance du point au village le plus proche

– SPOT-NDVI-100, SPOT-variance-100, SPOT-homogénéite-100, SPOT-dissimilarité-100 et SPOT-contraste-100 : les valeurs des descripteurs images SPOT calculés avec des polygones carrés de 100 m de coté

– la série temporelle (12 valeurs) de MODIS-NDVI, MODIS-variance, MODIS-homogénéite, MODIS-dissimilarité et MODIS-contraste : les valeurs des descripteurs images MODIS calculés avec des polygones carrés d'un kilomètre de côté pour chacun des mois concernés.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

dimension D_i	**intervalles de valeurs**				
id-pt	{1}, {2},. . . ,{498}				
date	{1}, {2},. . . ,{12}				
site-étude	{cinzana, koutiala, sikasso}				
type-utlisation-sol	{riz, sorgho, maïs,. . . }				
nom-village	{dioforongo, tigui, sanando,. . . }				
	proche	eloigne			
distance-village	[0,3000]	[3001,+∞[			
	très faible	faible	modéré	élevé	
SPOT-NDVI-100	[-1,0.2]	[0.2,0.3]	[0.3,0.5]	[0.5,1]	
	très faible	faible	modéré	élevé	trés élevé
SPOT-variance-100	$]-\infty, 0.56]$	$[0.56, 1.22]$	$[1.22, 2.38]$	$[2.38, 4.00]$	$[4.00, +\infty[$
SPOT-homogénéite-100	$]-\infty, 0.67]$	$[0.67, 0.74]$	$[0.74, 0.80]$	$[0.80, 0.87]$	$[0.87, +\infty[$
SPOT-dissimilarité-100	$]-\infty, 0.26]$	$[0.26, 0.40]$	$[0.40, 0.54]$	$[0.54, 0.72]$	$[0.72, +\infty[$
SPOT-contraste-100	$]-\infty, 0.27]$	$[0.27, 0.44]$	$[0.44, 0.72]$	$[0.72, 1.05]$	$[1.05, +\infty[$
MODIS-NDVI	$[-\infty, 2288]$	]2288, 2725]	]2725, 3497]	]3497, 4692]	]4692, $+\infty$[
Modis_contrast_1km	$]-\infty, 2.35]$	]2.35, 3.73]	]3.73, 5.35]	]5.35, 9.11]	]9.11, $+\infty$[
Modis_variance_1km	$]-\infty, 3.26]$	]3.26, 5.44]	]5.44, 8.15]	]8.15, 14.15]	]14.15, $+\infty$[
Modis_dissimilarite_1km	$]-\infty, 1.08]$	]1.08, 1.36]	]1.36, 1.66]	]1.66, 2.11]	]2.11, $+\infty$[
Modis_homogeneite_1km	$]-\infty, 0.37]$	]0.37, 0.45]	]0.45, 0.51]	]0.51, 0.57]	]0.57,$+\infty$[

Tableau 5. *Les dimensions D_i et le découpage de leurs domaines de valeurs $Dom(D_i)$ en intervalles.*

Les domaines $Dom(D_i)$ de chaque dimension D_i sont ensuite discrétisés en découpant leurs intervalles de valeurs. Pour l'indice de végétation NDVI, le découpage de l'intervalle de valeurs $[-1, 1]$ est donné par l'expert. Les domaines de valeurs des indices de texture sont découpés en cinq classes de même effectif. Le tableau 5 résume l'ensemble des dimensions D_i utilisées pour décrire les points de l'ensemble d'apprentissage et le découpage de leurs domaines de valeurs.

Une hiérarchie de valeur H_i est construite pour chaque dimension D_i. Les hiérarchies considérées sont toutes de profondeur 1 sauf pour l'attribut type-utlisation-sol (figure 7).

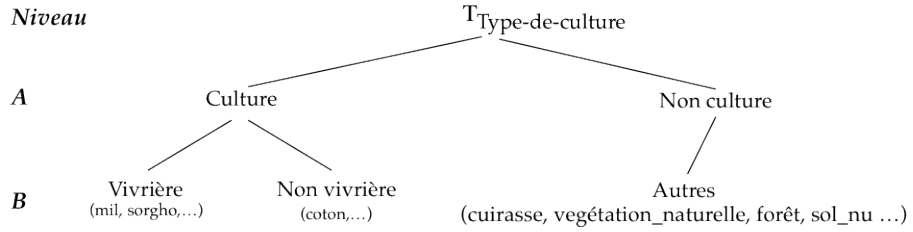

Figure 7. *La hiérarchie de valeurs H_i pour la dimension type-utlisation-sol.*

Dans l'étape de fouille, l'algorithme de recherche d'itemsets séquentiels fréquents M^3SP est appliqué sur les données d'apprentissage formatées, en choisissant un seuil

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

de support σ_{min}, un ensemble de dimensions de référence $\mathcal{D}_R$ et un ensemble de dimensions d'analyse $\mathcal{D}_A$.

Les itemsets séquentiels extraits sont présentés à l'analyste pour interprétation et validation.

5. Résultats et discussion

Dans cette section, nous analysons principalement les résultats obtenus à partir des images SPOT. Nous examinons ensuite les séquences obtenues sur les séries temporelles MODIS dans notre objectif de cartographie"culture" et "non culture", selon les valeurs associées à la dimension type-utilisation-sol représentées figure 7.

5.1. *Itemsets obtenus à partir des indicateurs sur SPOT*

Toutes les expériences ont été effectuées en prenant $\mathcal{D}_T = \{1\}$ et $\mathcal{D}_R = \{id - pt\}$. Comme la même estampille temporelle est associée à chaque point, les séquences obtenues ne sont constituées que d'un seul itemset. De plus, la dimension de référence étant l'identifiant des points, l'ensemble d'apprentissage est divisé en autant de blocs qu'il y a de points d'apprentissage. Par conséquent, dans ces expériences chaque itemset extrait est composé d'un seul item et son support correspond au nombre de points terrain qui le partagent.

Plusieurs expériences ont été menées en faisant varier les dimensions d'analyse $\mathcal{D}_A$ et en ne fouillant que les points situés sur un site d'étude donné (par filtrage de l'ensemble d'apprentissage avant la fouille).

Site d'étude	Itemset	Support
Cinzana (223 points)	$s_1 = \langle\{(culture, \top_{distance-village})\}\rangle$	138 (62%)
	$s_2 = \langle\{(culture, proche)\}\rangle$	121 (54%)
Koutiala (183 points)	$s_3 = \langle\{(culture, \top_{distance-village})\}\rangle$	105 (57%)
	$s_4 = \langle\{(culture, proche)\}\rangle$	80 (44%)
Sikasso (92 points)	$s_5 = \langle\{(culture, \top_{distance-village})\}\rangle$	46 (50%)
	$s_6 = \langle\{(culture, proche)\}\rangle$	27 (29%)

Tableau 6. *Itemsets extraits pour* $\mathcal{D}_A$={ *type-occupation-sol, distance-village* }.

Les itemsets présentés dans le tableau 6 ont été extraits en prenant comme dimensions d'analyse le type de culture et la distance au village le plus proche. Les résultats montrent que dans les 3 sites étudiés, les cultures sont généralement cultivées autour des villages, dans un rayon de 2 à 3 km pour la majorité. En effet, 88% des cultures de Cinzana ($121/138$, itemsets s_1 et s_2), 77% de celle de Koutiala ($80/105$, itemsets s_3 et s_4) et 59% de celles de Sikasso ($27/46$, itemsets s_5 et s_6) sont dans la couronne de 3 km autour des différents villages. Il n'apparaît pas possible de faire un lien

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

entre la distance au centre du village et le type de culture au vu du nombre trop faible de points terrain disponibles par site. Il a cependant été déjà observé dans plusieurs villages d'Afrique de l'Ouest un aménagement en auréoles. Le village et les jardins occupent une position centrale. Une premiére auréole (soforo) est constituée par les champs « de case » cultivés en rotation annuelle. Une seconde auréole (kongo foro) est formée par les champs de brousse (mil, sorgho, arachides, coton. . .). Enfin, la brousse (kongo) fournit les produits de la chasse, de la cueillette, le bois d'œuvre et de feu. La distance entre ces 3 auréoles varie entre les villages. L'analyse des motifs extraits qui ont été validés par la vérité terrain permet de valider les différents choix réalisés pour le pré-traitement des données.

Site d'étude	Itemset	Support
Cinzana (223 points)	$s_7 = \langle\{$(culture,très faible) $\}\rangle$	74 (33%)
Koutiala (183 points)	$s_8 = \langle\{$ (culture,modéré)$\}\rangle$	56 (31%)
	$s_9 = \langle\{$ (culture,faible) $\}\rangle$	33 (18%)
Sikasso (92 points)	$s_{10} = \langle\{$ (culture,faible) $\}\rangle$	25 (28%)
	$s_{11} = \langle\{$ (culture,modéré)$\}\rangle$	20 (22%)

Tableau 7. *Itemsets extraits pour* $\mathcal{D}_A = \{$ *type-utlisation-sol, SPOT-NDVI-100* $\}$.

Les itemsets présentés dans le tableau 7 ont été extraits en prenant comme dimensions d'analyse le type de culture et le descripteur NDVI calculé avec des polygones carrés de 100 m de côté. Nous pouvons signaler que le cycle de production végétale est particulier au Mali : les cultures étant pour la majorité des cultures pluviales, la croissance des plantes est étroitement liée à la pluviométrie (quantité et répartition). Les supports des itemsets séquentiels sont plus fréquemment faibles concernant le NDVI à Cinzana, qu'à Koutiala, qu'à Sikasso. Ceci reflète bien le gradient bioclimatique au Mali. Il pleut moins au Nord qu'au Sud, et donc les plantes ont une activité photosynthétique inférieure à Cinzana qu'à Sikasso, en moyenne. D'autre part, pour le NDVI du mois de novembre, les cultures sont déjà entièrement récoltées à Cinzana, et partiellement récoltées à Koutiala et à Sikasso. Ceci explique les 54% de culture avec un NDVI « très faible » à Cinzana ($74/138$, itemsets s_1 et s_7), contre 98% et 94% de cultures avec un NDVI « faible » ou « modéré » (itemsets s_3, s_5 et s_8 à s_{11}) à Koutiala et Sikasso respectivement.

Les itemsets présentés dans le tableau 8 ont été extraits en prenant comme dimensions d'analyse le type de culture et les quatre descripteurs de texture calculés avec des polygones carrés de 100 m de côté. La présence de quatre indices de texture induit une difficulté à interpréter la présence des itemsets fréquents. Si l'on analyse seulement le contraste, on observe dans 20 à 25% des cas, un contraste « élevé » pour Cinzana (itemset s_{12}), « faible » pour Koutiala (itemset s_{13}) et « très faible » pour Sikasso. Ce contraste qui diminue du Nord au Sud peut s'expliquer par une différence de densité d'arbres dans les champs cultivés. Il serait en effet plus commun de trouver des arbres comme le Balanzan, le Néré ou le Karitier dans des champs de la région de

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

Site d'étude	Itemset	Support
Cinzana (223 points)	$s_{12} = \langle\{(culture, \top_{variance-100}, \top_{homognit-100}, \top_{dissimilarit-100}, \textit{élevé})\}\rangle$	56 (25%)
Koutiala (183 points)	$s_{13} = \langle\{(culture, \top_{variance-100}, \top_{homognit-100}, \top_{dissimilarit-100}, faible)\}\rangle$	35 (19%)
Sikasso (92 points)	$s_{14} = \langle\{(culture, \top_{variance-100}, trslev, trsfaible, trsfaible)\}\rangle$	23 (25%)
	$s_{15} = \langle\{(culture, \top_{variance-100}, \top_{homognit-100}, \top_{dissimilarit-100}, faible)\}\rangle$	18 (20%)

Tableau 8. *Itemsets extraits pour* $\mathcal{D}_A = \{$ *type-utlisation-sol, SPOT-variance-100, SPOT-homogénéité-100, SPOT-dissimilarité-100, SPOT-contraste-100* $\}$.

Cinzana, qu'à Koutiala ou Sikasso, ce qui expliquerait les brusques changements de radiométrie, et donc un indice de contraste élevé.

À la suite de cette étude, il apparaîtrait intéressant de faire un lien entre la distance au centre du village et le type de culture. Il doit exister un lien entre les espèces cultivées et la distance au village, que nous essaierons de déterminer à partir de données terrain supplémentaires. D'autre part, le NDVI et la texture pourraient être mis en relation avec des images du mois de septembre ou d'octobre. C'est la période pendant laquelle les cultures sont dans des phases de croissance différentes suivant les régions, et entre elles également, puisque le maïs par exemple, est récolté, alors que les autres cultures céréalières restent sur pied pour encore un mois voire deux. Nous essaierons également d'utiliser pour l'extraction d'indices des polygones plus grands, pour un calcul de texture optimum (la faible taille des polygones ne permet pas de détecter beaucoup de motifs de texture répétés) et enfin, d'utiliser des séries temporelles d'images MODIS pour prendre en compte le cycle de croissance de chaque occupation du sol et ainsi mieux les différencier.

5.2. *Séquences obtenues à partir des indicateurs sur MODIS*

Toutes les expériences ont été effectuées en prenant $\mathcal{D}_T = \{date\}$ et $\mathcal{D}_R = \{id - pt\}$. Ayant différentes estampilles temporelles, nous obtenons des séquences d'itemsets. De façon identique à la section précédente, les supports correspondent au nombre de points terrain qui partagent la séquence obtenue.

Le tableau 9 indique la proportion de séquences extraites qui sont *discriminantes* c'est-à-dire qui apparaissent dans une classe mais qui n'apparaissent pas dans les autres classes considérées selon le niveau des valeurs dans la hiérarchie. Par exemple, le motif <(modis variance=-inf-7.585,modis contrast=-inf-6.245)> n'apparaît que dans la classe culture. La taille des motifs obtenus ne dépasse pas 4 items. Il est intéressant de constater que le nombre de motifs discriminants est le plus impor-

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

Niveau	Classe	Nb séquences discriminantes	Nb séquences total	Proportion
B	Vivrière	6	9	66.67%
	Non vivrière	12	12	100%
	Autres	13	16	81.25%
A	Culture	3	10	30%
	Non Culture	4	11	36.36%

Tableau 9. *Proportion de séquences discriminantes par classe avec* $minSupp = 0.5$

Motifs séquentiels	**Classe**
<(distance_village=eloigne,modis_ndvi=2182.6-2927.0) (distance_village=-eloigne,modis_ndvi=2927.0-3671.5)>	Culture
<(modis_ndvi=2182.6-2927.0) (modis_ndvi=2182.6-2927.0) (modis_ndvi=2927.0-3671.5)>	Non culture
<(modis_variance=-inf-7.585,modis_contrast=-inf-6.245) (modis_variance=-inf-7.585,modis_contrast=-inf-6.245)>	Culture
<(site=cinzana,distance_village=proche, modis_contrast=-inf-6.245392, modis_homogeneite=0.550197-0.628796) >	Vivrière
< (site=koutiala,modis_dissimilarite=1.852252-2.321403, modis_contrast=6.245392-11.962048) >	NonVivrière

Tableau 10. *Quelques séquences extraites à partir des séries temporelles d'indicateurs des images MODIS :* $\mathcal{D}_A = \{$ *distance_village, modis_ndvi, modis_variance, modis_homogénéité, modis_dissimilarité, modis_contrast* $\}$.

tant pour les valeurs vivrière et non vivrière. De plus, il apparaît difficile d'identifier à partir des indicateurs MODIS une réelle distinction entre les valeurs culture et non culture.

Le tableau 10 décrit quelques séquences obtenues selon les valeurs culture et non culture, vivrière et non vivrière. Les séquences obtenues selon les valeurs culture et non culture donnent des informations sur l'évolution du NDVI au cours du temps. Nous avions montré à travers la fouille sur les images SPOT que le NDVI reflétait le gradient bioclimatique au Mali. De par l'étude MODIS, nous pouvons suivre la phénologie des cultures et de la végétation naturelle à travers les motifs. Pour pou-

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

voir différencier par site, nous envisageons l'utilisation d'une base de données plus importante.

Nous pouvons souligner les deux séquences obtenues selon les valeurs « vivrière » et « non vivrière » qui impliquent les dimensions texturales. Tout comme pour SPOT, la distance au village inférieure à 3km apparaît sur le site de Cinzana. Concernant la texture, les deux sites de Koutiala et Cinzana présentent des motifs faisant apparaître la dimension texturale. Le motif de Cinzana possède un contraste moins élevé pour une culture vivrière que celui de Koutiala pour une culture non vivrière. Le motif de Cinzana présente également un item d'homogénéité, alors que celui de Koutiala présente un item de dissimilarité. Nous pouvons donc en conclure que la texture est différente suivant le type de culture considéré, mais également suivant le site d'étude. Cette prise en compte de la texture est donc très importante à des fins de classification.

Par ailleurs, l'analyse préliminaire des motifs a montré qu'il n'y a aucune variation temporelle suffisante de texture. Or, la texture devrait, tout comme le NDVI apparaître avec des valeurs différentes dans un motif. Nous émettons l'hypothèse que les intervalles de discrétisation choisis sont trop grands pour capter des variations temporelles de texture et recommandons une discrétisation plus fine de ces intervalles.

6. Conclusion

Nous avons présenté la première étape d'extraction d'itemsets et de séquences multidimensionnels d'une méthode de caractérisation des paysages ruraux et de leurs systèmes de culture. Nous avons mis en œuvre ce processus d'extraction sur des données du Mali et avons pu fouiller des données hétérogènes comme les relevés terrain avec des indicateurs obtenus à partir des des images satellitaires.

Cette approche préliminaire nous a permis :

– de retrouver des caractéristiques connues de l'agriculture ouest-africaine, telles que la structure du domaine cultivé autour d'un village sous forme d'auréoles, ou encore les différences temporelles de croissance de végétation entre un site très pluvieux et un autre plus sec ;

– mais aussi de révéler l'intérêt d'utiliser la texture d'images à moyenne résolution spatiale dans la discrimination des surfaces cultivées. En effet, jusqu'à présent et à notre connaissance, seule la texture d'images à haute et très haute résolution spatiale a été utilisée à cette fin. Dans cette étude, la taille de la fenêtre utilisée pour calculer la texture est de l'ordre de grandeur d'un terroir villageois. A cette échelle, la texture est donc dépendante de l'occupation du sol et de la fragmentation des paysages, qui sont des paramètres discriminants entre paysages agricoles et non agricoles.

Il s'agit à présent de poursuivre ces travaux (1) afin de prendre en compte la partie séquentielle des données (séquences extraites des images MODIS) qui seront le support du mécanisme de classification des types de culture puis (2) de proposer une évaluation (rappel, confiance) afin de mesurer l'efficacité d'une telle approche.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

7. Bibliographie

Agrawal R., Srikant R., « Mining Sequential Patterns », *in* P. S. Yu, A. L. P. Chen (eds), *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, IEEE Computer Society, p. 3-14, 1995.

Aksoy S., Koperski K., Tusk C., Marchisio G., « Interactive training of advanced classifiers for mining remote sensing image archives », *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, ACM, New York, NY, USA, p. 773–782, 2004.

dos Santos Silva M. P., Câmara G., de Souza R. C. M., Valeriano D. M., Escada M. I. S., « Mining Patterns of Change in Remote Sensing Image Databases », *ICDM*, p. 362-369, 2005.

Fayyad U. M., Piatetsky-Shapiro G., Smyth P., « From Data Mining to Knowledge Discovery : an Overview », *Advances in knowledge discovery and data mining*, vol. 1, p. 1–34, 1996.

Fritz S., See L., Rembold F., « Comparison of global and regional land cover maps with statistical information for the agricultural domain in Africa », *International Journal of Remote Sensing*, vol. 31, n° 9, p. 2237 - 2256, 2010.

Hansen M., Reed B., « A comparison of the IGBP DISCover and University of Maryland 1km global land cover products », *International Journal of Remote Sensing*, vol. 21, n° 6-7, p. 1365-1373, 2000.

Haralick R. M., « Statistical and Structural Approaches to Texture », *Proceedings of the IEEE*, vol. 67, n° 5, p. 786-804, 1979.

Haralick R. M., Shanmugam K., Dinstein I., « Textural Features for Image Classification », *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, n° 6, p. 610–621, November, 1973.

Julea. A., Meger N., Bolon P., « On mining pixel based evolution classes in satellite image time series », *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, p. 6, 2008.

Julea A., Méger N., Bolon P., Rigotti C., Doin M.-P., Lasserre C., Trouve E., Lazarescu V., « Unsupervised Spatiotemporal Mining of Satellite Image Time Series using Grouped Frequent Sequential Patterns », *IEEE Transactions on Geoscience and Remote Sensing*, 2011. To appear, vol. 49, issue 4, 2011, 14 pages.

Leenhardt D., Cernesson F., Mari J.-F., Mesmin D., « Anticiper l'assolement pour mieux gérer les ressources en eau : comment valoriser les données d'occupation du sol ? », *Ingénieries eau agriculture territoires*, vol. , n° 42, p. 13 - 22, June, 2005.

Marascu A., Masseglia F., « Mining sequential patterns from data streams : a centroid approach », *Journal of Intelligent Information Systems*, vol. 27, n° 3, p. 291-307, 2006.

Masseglia F., Poncelet P., Teisseire M., Marascu A., « Web usage mining : extracting unexpected periods from web logs », *Data Mining and Knowledge Discovery (DMKD)*, vol. 16, n° 1, p. 39-65, 2008.

Pei J., Han J., Mortazavi-Asl B., Zhu H., « Mining Access Patterns Efficiently from Web Logs », *in* T. Terano, H. Liu, A. L. P. Chen (eds), *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, April 18-20, 2000, Proceedings*, Lecture Notes in Computer Science, Springer, p. 396-407, 2000.

Author-produced version of the article published in Revue Internationale de Géomatique, 2011, 21(4), 469-488.
The original publication is available at http://geo.e-revues.com
http:// dx.doi.org/10.3166/rig.21.469-488

Perera D., Kay J., Koprinska I., Yacef K., Zaïane O. R., « Clustering and Sequential Pattern Mining of Online Collaborative Learning Data », *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, n° 6, p. 759–772, 2009.

Petitjean F., Gançarski P., Masseglia F., Forestier G., « Analysing Satellite Image Time Series by Means of Pattern Mining », *Intelligent Data Engineering and Automated Learning - IDEAL 2010, 11th International Conference, Paisley, UK, September 1-3, 2010. Proceedings*, vol. 6283 of *Lecture Notes in Computer Science*, Springer, p. 45-52, 2010.

Pinto H., Han J., Pei J., Wang K., Chen Q., Dayal U., « Multi-dimensional sequential pattern mining », *CIKM '01 : Proceedings of the tenth international conference on Information and knowledge management*, ACM, New York, NY, USA, p. 81–88, 2001.

Plantevit M., Laurent A., Laurent D., Teisseire M., Choong Y. W., « Mining multidimensional and multilevel sequential patterns », *ACM Transactions on Knowledge Discovery from Data TKDD*, 2010.

Rabatel J., Bringay S., Poncelet P., « Aide à la décision pour la maintenance ferroviaire préventive », *in* S. B. Yahia, J.-M. Petit (eds), *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, Cépaduès-Éditions, p. 363-368, 2010.

Romani L. A. S., de Ávila A. M. H., Jr. J. Z., Jr. C. T., Traina A. J. M., « Mining Relevant and Extreme Patterns on Climate Time Series with CLIPSMiner », *JIDM*, vol. 1, n° 2, p. 245-260, 2010.

Rouse I., « The explanation of culture change », *Science*, vol. 185, p. 343-344, 1974.

Salle P., Bringay S., Teisseire M., « Mining Discriminant Sequential Patterns for Aging Brain », *in* C. Combi, Y. Shahar, A. Abu-Hanna (eds), *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, p. 365-369, 2009.

Vintrou E., Desbrosse A., Bégué A., Traoré S., Baron C., LoSeen D., « Crop area mapping in West Africa using landscape stratification of MODIS time series and comparison with existing global land products », *Journal of Applied Earth Observation and Geoinformation - Under revision*, 2011.

Wang K., Xu Y., Yu J. X., « Scalable sequential pattern mining for biological sequences », *CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, USA, p. 178–187, 2004.