Fouille de données complexes : des relevés terrain aux données satellitaires pour la cartographie de paysages agricoles

Fadi Badra* — **Elodie Vintrou** ** — **Agnès Bégué**** — **Maguelonne Teisseire***

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

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

{*nom*}*@teledetection.fr*

RÉSUMÉ. *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 relevé terrain aux indicateurs de texture extraits des images satellites. Sa mise en œuvre sur des données associées au Mali pose les premières bases d'une méthode originale d'extraction de motifs séquentiels à partir de données complexes.*

ABSTRACT. *Remote sensing mining*

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

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

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) ou Data Mining (Fouille de Données). 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* » [FAY 96]. 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 prétraitement (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 déviation.

Notre objectif est de développer et de mettre en œuvre des méthodes de caractérisation des paysages ruraux et de leurs systèmes de culture, en utilisant des techniques de fouille de données permettant de coupler des données hétérogènes comme les relevés terrains à des informations issues des images satellites. Ce projet vise ainsi à proposer une approche alternative pertinente pour définir un mécanisme d'apprentissage basé sur des images satellitaires 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 critères d'implication qui n'auraient pas pu être identifiés autrement. Nous procédons suivant deux étapes :

– Adapter un mécanisme de recherche de motifs séquentiels multidimensionnels, comme proposé dans [PLA 10], à la fouille de séries d'images satellitaires et des don-

nées pouvant être mises en corrélation (informations externes) ;

– Proposer un système de catégorisation (classification supervisée) basé sur un mécanisme d'apprentissage afin d'offrir une aide à la cartographie de l'utilisation des sols. Cette solution doit être le moins sensible possible au changement d'échelle. Les informations annexes associées à des relevés terrains ou au croisement de bases de données (modèles numériques de terrain, climat...) seront donc très importantes.

Dans cet article, nous nous concentrons sur la première étape de recherche d'itemsets fréquents à partir de données de télédétection. 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-source, puisqu'elles résultent du croisement de données isssues de différents capteurs, de relevés terrain et bases de données externes. 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 [PIN 01], qui peuvent prendre en compte le caractère multidimensionnel des données.

Les résultats présentés dans cet article sont des résultats préliminaires mettant en œuvre un jeu de données restreint en termes d'imagerie satellitaire (images SPOT seulement) et de données exogènes. Cette approche préliminaire simplifiée nous a semblé indispensable pour assurer des bases communes aux deux communautés engagées dans ce travail, les agronomes et les spécialistes de fouille de données.

Dans les sections suivantes, nous décrivons tout d'abord les données étudiées (section 2) pour ensuite présenter les définitions associés aux motifs multidimensionnels (section 3). Puis, nous décrivons en détail le processus d'extraction mis en œuvre 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 Nord-Sud, 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).

Site d'étude	Précipitations moyennes	Cultures principales	Végétation naturelle (en majorité)	Pourcentage de surface cultivée
Cinzana (zone soudano-sahélienne)	600 mm/an	mil et sorgho	Végétation dégradée et sol nu	43%
Koutiala (zone soudano-sahélienne)	750 mm/an	mil, sorgho et coton	Végétation ouverte et fermée	52%
Sikasso (zone soudanienne)	1000 mm/an	maïs, coton et fruits	Végétation naturelle et dense	40%

Tableau 1. *Les principales caractéristiques des trois zones d'étude retenues.*

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 sahéliens. Au total, 980 points GPS ont été enregistrés, et des paysans de chacune des régions étudiées ont été interrogés. 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.

2.3. *Les données image*

Pour couvrir de larges zones avec précision, les images SPOT à 10 m de résolution constituent souvent la solution la plus efficace et la plus rentable. Pour cette étude, les images utilisées sont des images SPOT5 orthorectifiées et fusionnées, à 2.5 m en multispectral. Pour des raisons de disponibilité des images sur nos sites d'études, la date d'acquisition est le 14 novembre 2007 pour Cinzana et le 20 novembre 2007 pour Koutiala et Sikasso. Cependant, les classes d'occupation du sol observées on globalement peu évolué en deux ans (données terrain en 2009).

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. Ainsi, ces indices permettent de distinguer une surface de végétation verte photosynthétiquement active (indice élevé) avec une sur-

face de sol nu, ou faiblement couverte (indice faible). L'indice le plus souvent utilisé est le « *Normalized Difference Vegetation Index* » [ROU 74] :

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

L'indice varie de 0,1 (sol nu) à 0,9 (végétation verte et dense).

– *Les indices de texture.* L'hétérogénéité spatiale de la couverture végétale est étudiée au travers de la distribution spatiale des variables radiométriques. La variabilité spatiale d'une image est représentée par le concept de texture. Haralick élargit dans [HAR 79] 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.

En télédétection, l'utilisation de variables texturales dans les algorithmes de classification augmente la précision des résultats de classification par rapport à l'utilisation de la seule information spectrale [HAR 73]. Les images utilisées étant des champs continus de variables radiométriques, l'approche microtexturale est plus adaptée. 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 [HAR 73, HAR 79]. Quatorze indices (définis par Haralick) qui correspondent à des caractères descriptifs des textures peuvent être calculés à partir de cette matrice. Dans cette étude, les indices d'homogénéité, variance, dissimilarité, et contraste ont été calculés sur une fenêtre de 15×15 pixels.

3. Extraction d'itemsets séquentiels multidimensionnels

Le problème de la recherche de motifs séquentiels a été introduit par R. Agrawal dans [AGR 95] et appliqué avec succès dans de nombreux domaines comme la biologie [WAN 04, SAL 09], la fouille d'usage du Web [PEI 00, MAS 08], la détection d'anomalie [RAB 10], la fouille de flux de données [MAR 06] ou la description des comportements au sein d'un groupe [PER 09]. Des approches plus récentes [JUL 08] 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'aucun 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.

3.1. *Itemsets séquentiels multidimensionnels*

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.

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 particulière 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 1 présente un exemple de hiérarchie de valeurs H_i pour la dimension `site-étude`.

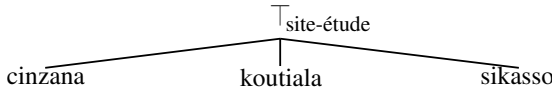

Figure 1. *La hiérarchie de valeurs H_i pour la dimension `site-étude`.*

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, $e = (\texttt{sorgho}, \texttt{cinzana})$ et $e' = (\texttt{sorgho}, \top_{\texttt{site-étude}})$ sont des items multidimensionnels qui décrivent des points terrain sur les dimensions d'analyse $\mathcal{D}_A = \{\texttt{type-de-culture}, \texttt{site-étude}\}$. 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 $(\texttt{sorgho}, \texttt{cinzana}) \subseteq (\texttt{sorgho}, \top_{\texttt{site-étude}})$ car `cinzana` est une spécialisation de $\top_{\texttt{site-étude}}$ dans la hiérarchie $H_{\texttt{site-étude}}$.

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 $e_i \not\supseteq e_j$). On définit une relation d'inclusion $\subseteq$ entre items multidimensionnels : un itemset i est inclus dans un itemset i' (noté $i \subseteq i'$) si pour chaque item a de i, il existe 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. On définit une relation de généralisation (ou spécialisation) entre séquences multidimensionnelles : une séquence $s = \langle i_1, i_2, \ldots, i_n \rangle$

pt-id	date	type-de-culture	site-étude	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-de-culture	site-étude	NDVI-100
2	1	mil	koutiala	faible
2	2	mil	koutiala	modéré

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

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 $\mathcal{D}_R = \{\texttt{type-de-culture}, \texttt{site-étude}\}$ est $(\texttt{mil}, \texttt{koutiala})$. 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}.

3.2. *Méthode*

L'algorithme M^3SP [PLA 10] effectue efficacement la recherche d'itemsets séquentiels multidimensionnels. Le choix de cet algorithme est motivé par notre objectif d'intégrer la dimension temporelle de séries d'images dans une prochaine étape.

4. Mise en œuvre

Un processus d'extraction de connaissances a été mis au point (figure 2). 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.

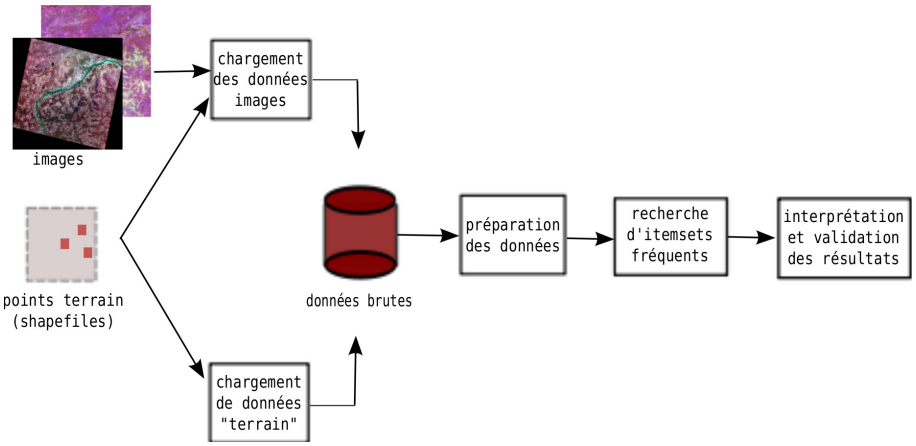

Figure 2. *Les différentes étapes du processus d'extraction de connaissances.*

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 : l'indice de végétation NDVI et les indices de texture variance, homogénéité, contraste et dissimilarité. À 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 un polygone carré centré sur ce point (figure 3). Deux jeux d'indices sont ainsi calculés avec des tailles de polygones de 100 m et 200 m de coté.

L'étape de préparation des données a pour but de constituer l'ensemble d'apprentissage et de formater les données pour 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 des relevés en zone de culture. On obtient 498 points dont la répartition par site d'étude et en culture / non culture est donnée par le tableau 4.

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.*

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

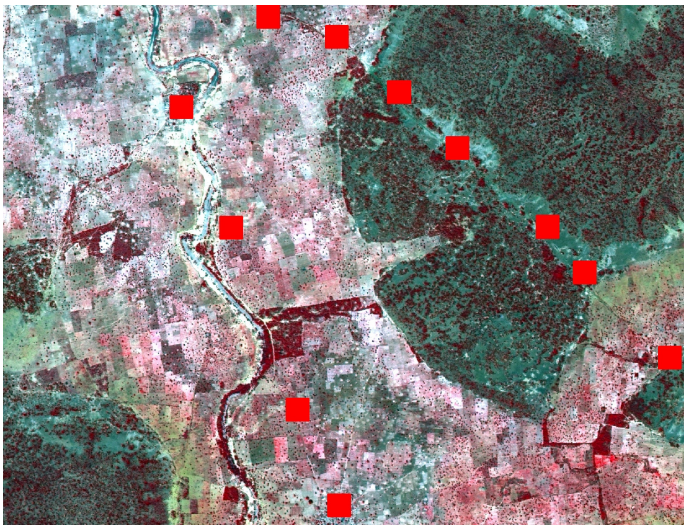

Figure 3. *Extraction des indices images pour chaque point terrain sur un polygone carré centré sur ce point (zoom sur une image SPOT du site de Cinzana, les carrés rouges sont des polygones de 100 m×100 m centrés sur chacun des points terrain). © CNES 2009, Distribution SPOT Image*

– `date` : estampille temporelle (constante dans cette expérience)

– `site-étude` : le nom du site d'étude

– `type-de-culture` : le type de 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

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

– `NDVI-200`, `variance-200`, `homogénéite-200`, `dissimilarité-200` et `contraste-200` : les valeurs des descripteurs images calculés avec des polygones carrés de 200 m de coté

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.

dimension D_i	intervalles de valeurs				
id-pt	{1}, {2},...,{498}				
date	{1}				
site-étude	{cinzana, koutiala, sikasso}				
type-de-culture	{riz, sorgho, maïs,...}				
nom-village	{dioforongo, tigui, sanando,...}				
	proche	eloigne			
distance-village	[0,3000]	[3001,+∞[			
	très faible	faible	modéré	élevé	
NDVI-100	[-1,0.2]	[0.2,0.3]	[0.3,0.5]	[0.5,1]	
NDVI-200	[-1,0.2]	[0.2,0.3]	[0.3,0.5]	[0.5,1]	
	très faible	faible	modéré	élevé	très élevé
variance-100 variance-200	$]-\infty, 0.56]$	$[0.56, 1.22]$	$[1.22, 2.38]$	$[2.38, 4.00]$	$[4.00, +\infty[$
homogénéite-100 homogénéite-200	$]-\infty, 0.67]$	$[0.67, 0.74]$	$[0.74, 0.80]$	$[0.80, 0.87]$	$[0.87, +\infty[$
dissimilarité-100 dissimilarité-200	$]-\infty, 0.26]$	$[0.26, 0.40]$	$[0.40, 0.54]$	$[0.54, 0.72]$	$[0.72, +\infty[$
contraste-100 contraste-200	$]-\infty, 0.27]$	$[0.27, 0.44]$	$[0.44, 0.72]$	$[0.72, 1.05]$	$[1.05, +\infty[$

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

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-de-culture` (figure 4).

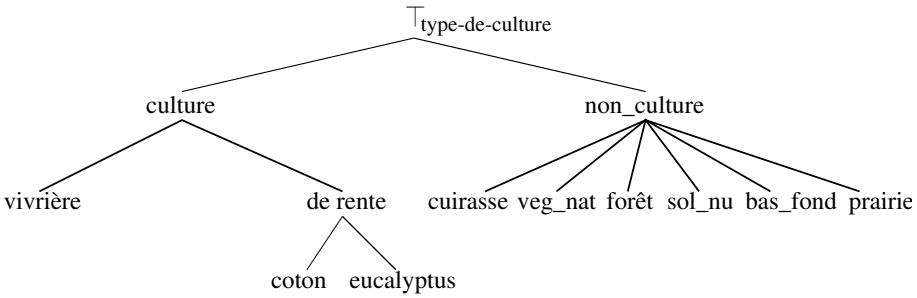

Figure 4. *La hiérarchie de valeurs H_i pour la dimension `type-de-culture`.*

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 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

Toutes les expériences ont été effectuées en prenant $\mathcal{D}_T = \{\texttt{date}\}$ et $\mathcal{D}_R = \{\texttt{id-pt}\}$. 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 résultat 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)	$s_1 = \langle\{(\texttt{culture}, \top_{\texttt{distance-village}})\}\rangle$	138 (62%)
	$s_2 = \langle\{(\texttt{culture}, \texttt{proche})\}\rangle$	121 (54%)
Koutiala (183 points)	$s_3 = \langle\{(\texttt{culture}, \top_{\texttt{distance-village}})\}\rangle$	105 (57%)
	$s_4 = \langle\{(\texttt{culture}, \texttt{proche})\}\rangle$	80 (44%)
Sikasso (92 points)	$s_5 = \langle\{(\texttt{culture}, \top_{\texttt{distance-village}})\}\rangle$	46 (50%)
	$s_6 = \langle\{(\texttt{culture}, \texttt{proche})\}\rangle$	27 (29%)

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

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 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 première 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.

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).

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

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

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 « medium » (itemsets s_3, s_5 et s_8 à s_{11}) à Koutiala et Sikasso respectivement.

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

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

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

Enfin, les différentes tailles de polygones n'induisent pas de changements notables dans les résultats.

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.

6. Conclusion

Nous avons présenté la première étape d'extraction d'itemsets 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 terrains avec des indicateurs obtenus des images satellitaires . Les premiers résultats sont prometteurs et laissent présager des perspectives d'analyse pertinentes. Il s'agit à présent de poursuivre ces travaux (1) afin de prendre en compte la partie séquentielle des données (séquences des images Modis) afin de trouver des motifs séquentiels 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.

7. Bibliographie

[AGR 95] AGRAWAL R., SRIKANT R., « Mining Sequential Patterns », YU P. S., CHEN A. L. P., Eds., *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, IEEE Computer Society, 1995, p. 3-14.

[FAY 96] 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, 1996, p. 1–34.

[HAR 73] HARALICK R. M., SHANMUGAM K., DINSTEIN I., « Textural Features for Image Classification », *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, n° 6, 1973, p. 610–621.

[HAR 79] HARALICK R. M., « Statistical and Structural Approaches to Texture », *Proceedings of the IEEE*, vol. 67, n° 5, 1979, p. 786-804, IEEE-Inst Electrical Electronics Engineers Inc.

[JUL 08] 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)*, 2008, page 6.

[MAR 06] MARASCU A., MASSEGLIA F., « Mining sequential patterns from data streams : a centroid approach », *Journal of Intelligent Information Systems*, vol. 27, n° 3, 2006, p. 291-307.

[MAS 08] 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, 2008, p. 39-65.

[PEI 00] PEI J., HAN J., MORTAZAVI-ASL B., ZHU H., « Mining Access Patterns Efficiently from Web Logs », TERANO T., LIU H., CHEN A. L. P., 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, 2000, p. 396-407.

[PER 09] 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, 2009, p. 759–772, IEEE Educational Activities Department.

[PIN 01] 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*, New York, NY, USA, 2001, ACM, p. 81–88.

[PLA 10] 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*, vol. 4, n° 1, 2010.

[RAB 10] RABATEL J., BRINGAY S., PONCELET P., « Aide à la décision pour la maintenance ferroviaire préventive », YAHIA S. B., PETIT J.-M., 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, 2010, p. 363-368.

[ROU 74] ROUSE I., « The explanation of culture change », *Science*, vol. 185, 1974, p. 343-344.

[SAL 09] SALLE P., BRINGAY S., TEISSEIRE M., « Mining Discriminant Sequential Patterns for Aging Brain », COMBI C., SHAHAR Y., ABU-HANNA A., 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, 2009, p. 365-369.

[WAN 04] 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*, New York, NY, USA, 2004, ACM, p. 178–187.