Mining sequential patterns from MODIS time series for cultivated area mapping

No Author Given

No Institute Given

Abstract. To predict and respond to famine and other forms of food insecurity, different early warning systems are using remote analyses of crop condition and agricultural production, using satellite-based information. To improve these predictions, a reliable estimation of the cultivated area at national scale must be carried out. In this study, we developed a datamining methodology for extracting cultivated domain patterns based on their temporal behavior as captured in time-series of moderate resolution remote sensing MODIS images.

Keywords: Knowledge Discovery, Data Mining, MODIS time series

1 Introduction

The northern fringe of sub-Saharan Africa is a region that is considered particularly vulnerable to climate variability and change, and food security remained there a major challenge. To address this issue, major international research efforts are being deployed, coordinated by the ongoing project AMMA (African Monsoon Multidisciplinary Analyses). Its aim is to better understand the West African Monsoon and its variability, and to improve the predictions of the impacts of this variability on West African societies.

One of the preliminary stages necessary for analyzing such impacts on agriculture and food security is a reliable estimation of the cultivated domain at national level, a scale compatible with climate change studies. For that purpose, different early warning systems such as FEWS and JRC-MARS use global land cover maps but they are generally focused on large ecosystems, and are not suitable for fragmented and heterogeneous African landscapes. Recent moderate-resolution sensors, such as MODIS/TERRA, with spatial resolutions as low as 250 m, offer new possibilities in the study of agricultural lands. With this increase in spatial resolution, the detection of groups of fields can now be considered. The low and medium spatial resolutions do not, by themselves, provide a completely satisfactory representation of the landscape but are compensated for by a large coverage area and by an excellent temporal resolution.

This brings us to the question whether moderate-resolution satellite data, in combination with external data (fields surveys, climate etc.) can provide a correct assessment of the distribution of the cultivated domain at country level. It is expected that more consistent information on vegetation would allow monitoring Sahelian rural landscapes with better continuity, thereby providing relevant

information for early warning systems.

In this study, we develop a datamining methodology to extract relevant sequential patterns to describe cultivated areas. These patterns are obtained from the static description and the temporal behavior as captured in time-series of moderate resolution remote sensing images. We applied this methodology in Mali, a representative country of the Sahel Belt of Africa. Both the temporal and spatial dimensions add substantial complexity to data mining tasks. A prioritization is needed to reduce the search space and to allow the relevant pattern extraction. We thus adopt a two-step approach: (1) identification of relevant descriptors per class (2) associated pattern mining from MODIS times series.

2 The Data Description

2.1 Study area

Mali is, after Senegal, the second westernmost country of West Africa around Latitude 14N. It displays a South North climatic gradient that ranges from subtropical to semi-arid, and which extends further north to arid and desertic. As for other West African countries along the same latitudinal belt, food security relies on an adequate supply of rainfall during monsoon season. This country can therefore be considered representative of the Sudano-Sahelian zone, where a strong dependence on rainfed agriculture implies vulnerability to major changes due to climate and human activities, and hence require specific attention. A particular attention was paid to 3 zones in Bani catchment, mainly located in Southern Mali (Table 1).

Table 1. Main characteristics of the three studied sites

Site name (eco-climatic zone)	Mean annual rainfall	Main crops	Natural vegetation type
Cinzana (Soudano-Sahelian)	600 mm	Millet, sorghum	High proportion of bare soils and sparse vegetation
Koutiala (Soudano-Sahelian)	750 mm	Cotton, millet, sorghum	Large areas of semi-open and closed natural vegetation
Sikasso (Soudanian)	1000 mm	Maize, cotton, fruit crops	Dense natural vegetation

2.2 Data

Field data Fields surveys were conducted in Mali during the 2009 and 2010 cropping seasons (from May to November) in order to characterize Soudano-Sahelian rural landscapes. Three sites (Cinzana, Koutiala, Sikasso) were selected to sample the main agro-climatic regions of Central and Southern Mali (Table 1). 980 GPS waypoints were registered, and farmers were interviewed. Each waypoint was transformed into a polygon whose center has been affected a land use.

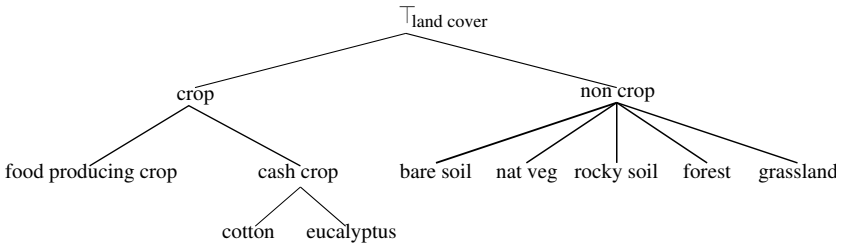

Fig. 1. Crop hierarchy

External data Six static descriptors were also used to characterize the site surveys:

- soil type
- distance to the village
- distance to the river
- rainfall
- ethnic group
- name of the village

Images data MODIS time series: The NASA Land Process Distributed Active Archive Center (LP DAAC) is the repository for all MODIS data. Amongst MODIS products, we selected the Vegetation Indices 16-Day L3 Global 250 m SIN Grid temporal syntheses for our study. For Mali, a set of 12 MODIS 16-days composite normalized difference vegetation index (NDVI) images (MOD13Q1/V05 product) at a resolution of 231.6 m were acquired for 2007 (we keep the best quality composite image out of two for each month). The year 2007 was chosen to overlap with the more recent high resolution data available. We assume that the observed classes of land use remained globally unchanged from 2007 to 2009 (fields surveys in 2009). However, Malian farmers pratice crop rotation. It is the practice of growing a series of dissimilar types of crops in the same area in sequential seasons for various benefits such as to avoid the build up of pathogens

and pests that often occurs when one species is continuously cropped, improving soil structure and fertility. Thus, we decided to only consider the two higher levels of the crop hierarchy (Figure 1).

Remotely-sensed indices used

- **Normalized Difference Vegetation Index:** NDVI is one of the most successful index to simply and quickly identify vegetated areas and their "condition", providing a crude estimate of vegetation health. It displays the relationship between the quantity of chlorophyll in leaves with red and near infrared wavelength, so that NDVI image is used to search vegetation as estimating biomass, plant productivity, fractional vegetation cover [10].

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

  where RED and NIR stand for the spectral reflectance measurements acquired in the red and near-infrared regions, respectively. In general, NDVI values range from -1.0 to 1.0, with negative values indicating clouds and water, positive values near zero indicating bare soil, and higher positive values of NDVI ranging from sparse vegetation (0.1 - 0.5) to dense green vegetation (0.6 and above). Furthermore, different land covers exhibit distinctive seasonal patterns of NDVI variation. Crops have generally a distinct growing season and period of peak greenness, which allows the discrimination with other types of land cover.
- **Texture:** Information content in a digital image is expressed by the intensity of each pixel (*i.e.,*. tone or color) and by the spatial arrangement of pixels (*i.e.,*. texture, shape, and context) in the image. Traditionally, tone (*i.e.,*. spectral intensity) has been the primary focus for most image analysis and hence information extraction in remote-sensing studies. However, texture analysis is examined as an important contributor to scene information extraction. The majority of image classification procedures, particularly in operational use, rely on spectral intensity characteristics alone and thus are oblivious to the spatial information content of the image. Textural algorithms, however, attempt to measure image texture by quantifying the distinctive spatial and spectral relationships between neighboring pixels. In response to the need for extracting information based on the spatial arrangement of digital image data, numerous texture algorithms have been developed. Statistical approaches, such as those developed by [4] make use of gray-level probability density functions, which generally are computed as the conditional joint probability of pairs of pixel gray levels in a local area of the image. In this study, we used four Haralick textural indices [3] calculated on the MODIS time series : variance, homogeneity, contrast and dissimilarity on ENVI. The Haralick textural features describe the spatial distribution of gray values and the frequency of one gray tone appearing with another gray tone in a specified distance and at a specified angle. The generation of these indices is based on different orientations of pixels pairs, with specific angle

(horizontal, diagonal, vertical, co-diagonal) and distance, called patterns. We determined empirically a size of pattern of 15 pixels for MODIS, which is the smaller patch repeated in different direction and distance.

3 Motivating Example

In order to illustrate our approach, we consider the following example that will be used throughout the paper as a running example. Let us consider a relational table T in which $NDVI$ values by field are stored. More precisely, we assume that T is defined over six dimensions (or attributes) as shown in Table 2 and where: D is the date of statements (considering two dates, denoted by 1 and 2), I is the field identifier (considering 4 different fields, denoted by F1,F2, F3 and F4), C is the crop type (considering two discretized values, denoted by FP (food-producing) and NFP (non food-procucing)), S is the soil type (considering three different soil types, denoted by GS (gravelly soils), SL (sandy loam) and CL (clay loam)), DV is the distance between the associated field and the nearest village (considering two discretized values, denoted by near and far), $NDVI$ stands for the $NDVI$ value associated to each field at each timestamp (considering 4 abstract values n_1, n_2, n_3 and n_4). We consider five sets of dimensions as follows: *(i)* the dimension D representing the date, *(ii)* the dimension I representing the identifier, *(iii)* the dimensions S and DV, that we call *static dimensions* or *descriptors* (values of these dimensions associated to a given field do not evolve over time), *(iv)* the dimension $NDVI$, that we call *dynamic dimension* or *indicators* (values of these dimensions associated to a given field evolve over time) and *(v)* the dimension C that we call the class. For instance, the first tuple of T (Table 2) means that the field 1 is a food-producer crop composed by CL, near to a village and that , at date 1, the NDVI value was n_1. Observing in great details the static attribute values per class, some comments should be made. First, food-producing crops are always located near to the village whereas the soil composition is changing. Similarly, non food-producing crops are always cultivated on GS whereas the distance to the nearest village is changing. A first interpretation to these comments is that the dimension DV appears to be decisive to identify food-producing crops whereas the dimension S appears to be decisive to identify non food-producing crops. Consequently, it is pertinent to only consider decisive dimensions per crop to mine representative rules. Once static dimensions have been filtered, the dynamic dimension (NDVI) is considered in order to mine sequential patterns characterizing crops. Let us suppose that we look for sequences which are verified by all the crops in a given class . Under this condition, the pattern $\langle(near, n_1)(near, n_2)\rangle$ (meaning that fields located near to a village and where the NDVI statement are n_1 at a certain date and n_2 after) characterizes the food-producing crops and the pattern $\langle(GS, n_3)\rangle$ characterizes the non food-producing crops. It should be noted that representative rules per class can be composed by values of different dimensions. In the rest of this paper, we describe the adopted methodology to determine the deci-

sive attributes per class and how the table T is subdivided and mined to obtain representative rules per class.

Table 2. Table T

D (Date)	I (Id)	C (Crop)	S (Soil)	D (Distance to village)	NDVI (NDVI value)
1	F1	FP	CL	near	n_1
1	F2	FP	SL	near	n_1
1	F3	NFP	GS	far	n_2
1	F4	NFP	GS	near	n_3
2	F1	FP	CL	near	n_2
2	F2	FP	SL	near	n_2
2	F3	NFP	GS	far	n_4
2	F4	NFP	GS	near	n_3

4 Preliminary Definitions

In this section, concepts and definitions concerning multidimensional sequential patterns are presented and are inspired by the notations introduced in [9]. For each table defined on the set of dimensions D, we consider a partition of D into three sets: D_t for the temporal dimension, $D_{\mathcal{A}}$ for the analysis dimensions and D_R for the reference dimension. Each tuple $c = (d_1, \ldots, d_n)$ can thus be denoted $c = (r, a, t)$ with r the restriction on D_R, a the restriction on D_A and t the restriction on D_t .

Definition 1. *(Multidimensional Item) A multidimensional item e defined on $D_{\mathcal{A}} = \{D_{i1}, \ldots, D_{im}\}$ is a tuple $e = (d_{i1}, ..., d_{im})$ such that $\forall k \in [1, m], d_{ik} \in Dom(D_{ik})$.*

Definition 2. *(Multidimensional Sequence) A multidimensional sequence S defined on $D_{\mathcal{A}} = \{D_{i1}, \ldots, D_{im}\}$ is an ordered non empty list of multidimensional items $S = \langle e_1, \ldots, e_l \rangle$ where $\forall j \in [1, l]$, e_j is a multidimensional item defined on $D_{\mathcal{A}}$.*

Considering our running example and that $D_{\mathcal{A}} = \{DV, NDVI\}$, $(near, n_1)$ is a multidimensional item. $\langle (near, n_1)(near, n_2) \rangle$ is a multidimensional sequence on $D_{\mathcal{A}}$.

Remark 1. In the original framework of sequential patterns [1], a sequence is defined as an ordered non empty list of itemsets where an itemset is a non empty set of item. Nevertheless, in the scope of this paper, we only consider sequences of item since at each date, one and only one item can occur for each field. For instance, only one NDVI statement is available per date and field.

An identifier is said to support a sequence if a set of tuples containing the itemsets satisfying the temporal constraints can be found.

Definition 3. *An identifier $r \in Dom(D_R)$ supports a sequence $S = \langle e_1, \ldots, e_l \rangle$ if $\forall j \in 1 \ldots l$, $\exists d_j \in Dom(D_t)$, $\exists t = (r, e_j, d_j) \in T$ where $d_1 < d_2 < \ldots < d_l$.*

Definition 4. *Sequence Support Let D_R be the reference dimension and T the table. The support of a sequence S is:*

$$support(S) = \frac{|\{r \in Dom(D_R) \text{ s.t. } r \text{ supports } S\}|}{|Dom(D_R)|}$$

Definition 5. *(Frequent Sequence) Let $minSupp \in [0, 1]$ be the minimum user-dened support value. A sequence S is said to be frequent if $support(S) \geq minSupp$.*

Considering the definitions above, an item can only be retrieved if there exists a frequent tuple of values from domains of $D_{\mathcal{A}}$ containing it. For instance, it can occur that neither $(CL, near)$ nor $(SL, near)$ nor $(GS, near)$ is frequent whereas the value *near* is frequent. Thus, [9] introduces the *joker* value $*$. In this case, we consider $(*, near)$ which is said to be *jokerized*.

Definition 6. *Jokerized Item Let $e = (d_1, \ldots, d_m)$ a multidimensional item. We denote by $e_{[d_i/\delta]}$ the replacement in e of d_i by δ. e is said to be a* jokerized *multidimensional item if:* (i) $\forall i \in [1, m], d_i \in Dom(D_i) \cup \{*\}$, (ii) $\exists i \in [1, m]$*such that $d_i \neq *$ and* (iii) $\forall d_i = *, \nexists \delta \in Dom(D_i)$ *such that $e_{[d_i/\delta]}$ is frequent.*

A *jokerized* item contains at least one specified analysis dimension. It contains a $*$ only if no specific value from the domain can be set. A *jokerized* sequence is a sequence containing at least one *jokerized* item.

5 Method

5.1 Overview

In this paper, we aim at discovering representative rules in order to caracterize crop classes and propose a four-step method to achieve this issue. It should be noticed that the crop classes depends on the user-defined interest level of the crop hierarchy displayed in Fig 1. For instance, assuming that the user would like to discover representatives rules for classes in the second level of the hierarchy, the set of classes will be {*food-producing,non food-producing, other*}. These four steps are illustrated in Fig. 2 and are briefly presented here:

1. **The raw database pretreatment.** During this phase, two actions are performed. First, since the raw database stores crops at the lowest level of the hierarchy, these attributes values must be rewritten to match with the user-defined interest level. Second, sequential pattern mining aims at discovering frequent relations in a database but is not well adapted to mine numerical attributs (*e.g.,* distance to the village, NDVI value) due to the huge definition domain of such attributes. Consequently, numerical attributes are discretized to improve the sequential pattern mining phase.

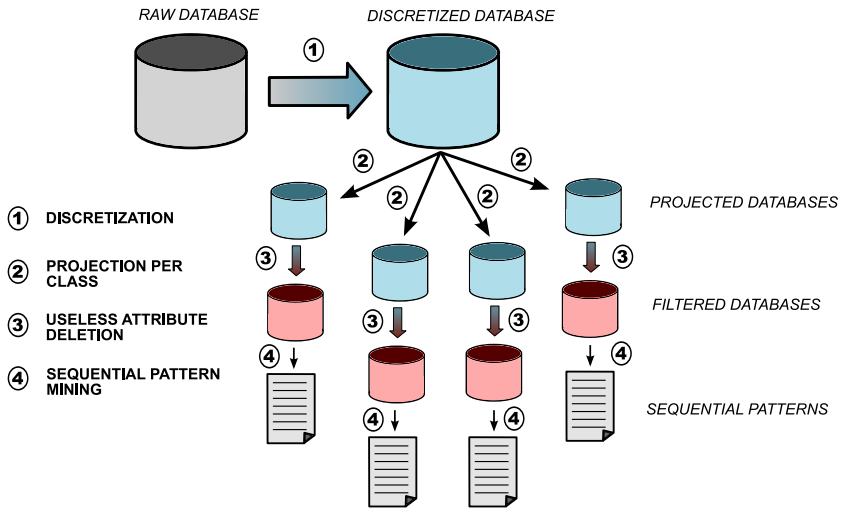

Fig. 2. Overall schema of the proposed methodology

2. **The build of projected databases.** Since we would like to obtain representative rules per class, the pretreated database is projected on the different class values.
3. **The decisive attribute computation.** During this step, a search is performed on each projected databases in order to find and delete non-decisive static attributes dimensions. Intuitively, a static attribute is said to be non-decisive if none of its value allows to characterize the class. More precisely, we guarantee that if it does not exist any value of a static attribute appearing in at least $minSupp\%$ in the projected database, the representative rules associated to this class wont never contain specific values of this static attribute. Consequently, it is useless to consider it in the rest of the process and this attribute will be removed from the projected database
4. **The sequential pattern mining.** Once the projected databases were cleaned up, the algorithm M^2SP is applied on each databases. We obtain a set of frequent pattern for each class.

Theses steps are now detailed in the following subsections.

5.2 The Database Pretreatment and Projections

The first performed treatment is the rewriting of the database in order to make the crop attribute values and the user-defined interest level match. This is motivated by two reasons. First, as mentioned in Section 2, mining representative rules for precise crop values is not consistent. As a consequence, crop attribute values must be rewritten to, at least, the above level of granularity. Second,

since the hierarchy is composed by two workable levels of granularity, it is interesting to allow user to choose which level must be explored. Consequently, an user-defined parameter, *Level*, is introduced to specify which level of granularity to mine. Thus, rules representing different generalized classes can be compared. Ann illustration of this database rewriting is displayed in Table 2 where crop attribute values have been already generalized to the second level of granularity (i.*e.,* $Dom(Crop) = \{FP, NFP\}$).

A second pretreatment is the discretization of numerical attributes. This discretization is motivated by the use of the sequential pattern technique to mine representative rules. Indeed, sequential pattern algorithms aim at discovering frequent relation among the fields belonging to the same class. When dealing with numerical attributes, two values can be considered as different items even if they are very close. For instance, let us consider that the distance to the nearest village is 200 meters for field 1 and is 205 meters for the field 2. These two distances woud have been considered as different items by the M^2SP algorithm without discretization even if they are semantically closed. In our application case, numerous attributes are numerical. Thus, this discretization is necessary. Numerous discretization techniques can be found in the literature [2]. Section 6 details the adopted technique per numerical attribute.

Once the database was pretreated, projection per crop attribute values is performed. Indeed, this is motivated by the fact that we would like to discover representative rules per class. Thus, an intuitive way to achieve this goal is to subdivide the pretreated table into smaller ones associated to each class. Regarding our running example, Tables 3 and 4 display the result of this projection.

Table 3. T_{FP}, the FP projected table

D (Date)	I (Id)	S (Soil)	D (Distance to village)	NDVI (NDVI indice)
1	F1	CL	near	n_1
1	F2	SL	near	n_1
2	F1	CL	near	n_2
2	F2	SL	near	n_2

5.3 Dimensionality Reduction

Once the projected databases were built, a search is performed on the static attributes of each database in order to identify useless static attributes. Intuitively, if values of a static attribute are very changing, this attribute is not really characteristic to this class. So, it can be deleted from the projected class. The main advantage of such a strategy is to reduce the search space during the sequential pattern mining phase. Indeed, it is empirically shown in [9] that the number of dimensions exponentially impacts on both the memory consumption and the

Table 4. T_{NFP}, the NFP projected table

D (Date)	I (Id)	S (Soil)	D (Distance to village)	NDVI (NDVI indice)
1	F3	GS	far	n_2
1	F4	GS	near	n_3
2	F3	GS	far	n_4
2	F4	GS	near	n_3

extraction time. Whereas traditional applications domains often deal with few analysis dimensions, this point can be very problematic in our context since the number of both static and dynamic dimensions can be high. For instance, experiment results presented in Section 6 concern at most 12 dimensions. Traditional multidimensional sequential pattern approaches cannot efficiently deal with such a number of analysis dimensions. Moreover, independently to performance considerations it is important to notice that the higher the number of dimensions, the higher the number of extracted patterns. Since these extracted patterns will be exploited by experts, reducing the dimensionality without loss of expressivity is very relevant to improve the result analysis phase.

To perform such a dimensionality reduction, we proceed as follows. Let $minSupp$ be the user-defined parameter used during the sequential pattern mining phase, T_i be a projected database and $D_j \in T_i$ be one the static dimension in T_i. It can be easily proved that if it does not exist any value of D_j appearing in at least $minSupp \times |T_i|$ tuples in T_i (where $|T_i|$ is the size of T_i), it cannot exist any sequential pattern extracted from T_i where a value of D_j appears. If so, the dimension D_j is considered as useless and is thus deleted from T_i. A direct corollary of this propriety is that if an attribute is retained, it will exist at least one sequential pattern containing a value of D_j. To illustrate this affirmation, let us consider, T_{FP}, the projected database presented Table 3 and $minSupp = 1$. The two static attributes are D and S. Regarding the D attribute, all the tuples share the same value ($near$). This attribute is considered as useful for the next step and is thus retained. Let us now consider the S attribute. Here, no value satisfies the $minSupp$ condition. As a consequence, S is deleted from this table. To attest the consequence of such a strategy, let us consider, SP_{FP}, the set of the multidimensional sequential patterns extracted from T_{FP} where $minSupp = 1$, $D_t = D$, $D_R = I$ and $D_{\mathcal{A}} = \{C, S, NDVI\}$ (*i.e.,* all the static and dynamic attributes are considered). Under these conditions, $SP_{FP} = \{\langle(*, near, n_1)\rangle, \langle(*, near, n_1)(*, near, n_2)\rangle\}$. It is readily noticeable that D occurs in SP_{FP} but not S.

It is interesting to observe that the set of useful attributes per class can be different. As a consequence, independently to the values of these attributes, attributes themselves can be representative to one class. For instance, performing the above described dimensionality reduction technique on T_{NFP} (see Table 4), S but not D will be retained this time.

5.4 Mining Representative Rules

Once useless attributes have been deleted, the M^2SP algorithm is applied on each projected and cleaned database T_i such that $minSupp$ is the same as defined during the previous step, $D_t = D$, $D_R = I$ and $D_{\mathcal{A}}$ is composed by the the retained static attributes and the dynamic attributes. We note SP_{T_i} the set of sequential patterns extracted from T_i. For instance considering T_{FP} and $minSupp = 1$, $\langle (near, n_1)(near, n_2) \rangle$ is a frequent sequence meaning that NDVI values equals n_1 and then n_2 is a frequent behaviour for fields cultivating food-producing crops located near to a village.

6 Experiment Study

In this section, we present experiments to evaluate the feasability and efficiency of our approach. Throughout the experiments, we answer the following questions inherent to efficiency issues: *Does the dimensionality reduction technique allow to delete useless static attributes without loss of information? Does the mining process allow to discover discriminating patterns per class ? Does the texture data allow a better discriminating pattern extraction than only considering NDVI values?* The experiments were performed on a Intel(R) Xeon(R) CPU E5450 @ 3.00GHz with 2GB of main memory, running Ubuntu 9.04. The methods were written in Java 1.6. We first describe the adopted protocol and then present and discuss our results.

6.1 Protocol

The method was evaluated on the dataset described in Section 2. This dataset contains 980 distinct fields and a MODIS time serie of length 12 is associated to each field. The 7 static dimensions and the 5 dynamic dimensions were the same as described in Section 2. As mentioned in Section 5, a discretization step is necessary to efficiently mine frequent patterns. The adopted discretization methods are as follows:

- EQUI-WIDTH technique (the generated intervals have the same width) was used for distance village and distance river attributes
- EQUI-DEPTH technique (the generated intervals have the same size) was used for the other numerical attributes

In this experiment study, two sets of classes were considered. The first set of classes, denoted by B aims at discovering patterns allowing the distinction between food-producing crops (FP), non food-producing crops (NFP) and non crops (OTHER). The second set of classes, denoted by C, aims at discovering patterns allowing the distinction of more general classes : crops (Cr) and non crops (NCr).

In order to evaluate the impact of texture data in discriminating pattern extraction, we consider a first configuration, denoted by $Default$, where all the

dynamic attributes were used. On the contrary, the configuration denoted by $NDVI$ is only composed by NDVI values as dynamic attribute.

Three experiment results are presented and discussed in this section:

1. A first experiment was performed to evaluate the number of retained static attributes according to two *minSupp* values
2. A second experiment was performed to evaluate the number of discriminating patterns. Here, *discriminating* means that a pattern appears in one class but not in the others.
3. Finally, the last experiment was performed to observe the discriminating dimension values according the two above described configurations.

6.2 Results and Discussion

Figure 3 displays the the retained attributes according to the two sets of classes and two *minSupp* values. First of all, it can be noticed that the *minSupp* threshold value has an obvious impact on this attribute selection. Indeed, considering $minSupp = 0.5$, more than half of the attributes were deleted. Moreover, it is interesting to observe that the retained attributes per class and set of classes are roughly identical.

Fig. 3. Retained static attributes under default configuration (left: $minSupp = 0.5$ / right: $minSupp = 0.3$)

Level	Class	Static attributes						
		Distance village	Site name	Ethnic group	Rainfall	Soil type	Distance river	Village
B	FP	x		x	x			
	NFP	x		x				
	OTHER	x		x				
C	Cr	x		x				
	NCr	x		x				

Level	Class	Static attributes						
		Distance village	Site name	Ethnic group	Rainfall	Soil type	Distance river	Village
B	FP	x	x	x	x	x		
	NFP	x	x	x	x	x		
	OTHER	x	x	x	x	x		
C	Cr	x	x	x	x	x		
	NCr	x	x	x	x	x		

Figure 4 displays the proportion of discriminating patterns per class with $minSupp = 0.5$ and the NDVI configuration. Indeed, even if a pattern was extracted from one class, it is not enough to consider it as discriminating (*i.e.,* the same pattern can appear in different classes) Thus, queries was formulated to search which patterns appear in one class and not in the others. Two conclusions

can be drawn from this figure. First, considering the set of classes B, most of the extracted patterns are discriminating (even if the FP class obtains a worse score). Second, finding discriminating patterns on the set of classes C looks more difficult.

Fig. 4. Proportion of discriminating patterns per class with $minSupp = 0.5$ and the NDVI configuration

Level	Class	#disc. patterns	#patterns	Proportion
B	FP	6	9	66.67%
	NFP	12	12	100%
	OTHER	13	16	81.25%
C	Cr	3	10	30%
	NCr	4	11	36.36%

Figure 5 displays some representative discriminating attribute values according the two configurations and the two sets of classes. An attribute value is said to be discriminating if it does not appear in any pattern of the other classes. This experiment aims at observing the impact of texture dynamic values on the extracted patterns. Some conclusions can be drawn. First of all, the class $OTHER$ does not contain discriminating value independently to the configuration. Second, a very interesting and promising result is that the default configuration contains much more discriminating values than the NDVI configuration. Moreover, these discriminating values concern the texture attributes. This result reinforces our idea that texture attributes are very useful in automatic landscape recognition.

To conclude this experiment study, we have empirically shown that (1) the dimensionality reduction method allows to reduce the search space by deleting useless attributes. (2) Most of the extracted patterns are discriminating. (3) It appears to be more difficult to distinguish between Cr and NCr classes than FP, NFP and $OTHER$ classes with our approach. And (4), most of the discriminating attribute values concern the texture attributes.

7 Related Work

Applications of sequential pattern mining methods to Satellite Image Time Series (SITS) include [7, 5, 8, 6]. Interest in these methods to study change detection on satellital images come from the fact that they are (i) multi-temporal, (ii) robust to noise, (iii) able to handle large volumes of data, and (iv) capable of capturing local evolutions without the need for prior clustering.

In [5], sequential pattern mining is applied to study change in land cover over a 10 months period on a rural area of east Romania. Pattern extraction is

Fig. 5. Some discriminating dimension values per class with $minSupp = 0.3$ (top: default config. / bottom: NDVI config.)

Level	Class	Attribute	Value
B	FP	modis homogeneity 1km	0.48-0.52
		modis variance 1km	3.27-4.34
		modis dissimilarity 1km	1.36-1.51
	NFP	distance village	3150-6205
		modis contrast 1km	5.35-6.54
		modis dissimilarity 1km	1.51-1.66
	OTHER	NONE	
C	Cr	modis dissimilarity 1km	1.21-1.36
		modis variance 1km	3.27-4.34
	NCr	modis variance 1km	10.28-14.15

Level	Class	Attribute	Value
B	FP	NONE	
	NFP	rainfall	800
		distance village	3149.3-6205.6
	OTHER	NONE	
C	Cr	NONE	
	NCr	distance village	3149.3-6205.6

used to group together SPOT pixels that share the same spectral evolution over time. The SITS data is thus processed at the pixel level, by taking the values of the pixels on each of the SPOT bands. A method is proposed to visualize the extracted patterns on a single image.

[8] presents a similar approach but pixel values are computed from four SPOT bands instead of a single band. The SITS period coverage is also much longer: a 20-year time image series is mined in order to study urban growth in the south west of France. A visualization technique is proposed to locate areas of evolution. Results show that mining all pixels of the images leads to the generation of a huge number of non-evolution patterns. Additional strategies are then required to filter out all non informative patterns.

To the best of our knowledge, sequential pattern mining has only been applied at the pixel level on high resolution images without taking into account external data or texture information in the mining process. In this paper, we have shown that sequential pattern mining can help to characterize cultivated areas from moderate resolution remote sensing images MODIS.

8 Conclusion

The objective of this study was to propose an original method to extract sets of relevant sequential patterns from MODIS times series that can be used for

cultivated area mapping. We have developed a data mining method based on two steps and applied it in Mali. Experiment study conducted on this data set reinforce our intuition about the importance of texture attributes to improve the automatic landscape recognition. Our future work will be aimed at validating the extracted patterns per class. After which, we can go a step further to build the classifier based on these patterns and evaluate the predictions of the cultivated area at national scale.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.
2. J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Machine LearningEWSL-91*, pages 164–178. Springer, 1991.
3. R. Haralick. Statistical and structural approaches to texture(image type analysis). In *IEEE, Proceedings*, volume 67, pages 786–804, 1979.
4. R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man and cybernetics*, 3(6):610–621, 1973.
5. A. Julea., N. Meger, and P. Bolon. On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, page 6, 2008.
6. A. Julea, N. Méger, P. Bolon, C. Rigotti, M.-P. Doin, C. Lasserre, E. Trouve, and V. Lazarescu. 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.
7. A. Julea, N. Méger, and E. Trouvé. Sequential patterns extraction in multitemporal satellite images. In *17th European Conference on Machine Learning and 10th European Conference on Principles and Practice of Knowledge Discovery (ECML/PKDD 2006) - Berlin (Germany) The 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2006*, pages 94–97, Berlin Allemagne, 09 2006. -.
8. F. Petitjean, P. Gançarski, F. Masseglia, and G. Forestier. Analysing satellite image time series by means of pattern mining. In *Intelligent Data Engineering and Automated Learning - IDEAL 2010, 11th International Conference, Paisley, UK, September 1-3, 2010. Proceedings*, volume 6283 of *Lecture Notes in Computer Science*, pages 45–52. Springer, 2010.
9. M. Plantevit, Y. Choong, A. Laurent, D. Laurent, and M. Teisseire. M 2 SP: Mining sequential patterns among several dimensions. *Knowledge Discovery in Databases: PKDD 2005*, pages 205–216, 2005.
10. I. Rouse. The explanation of culture change. *Science*, 185:343–344, 1974.