A DATA MINING APPROACH TO DISCOVER COLLECTIONS OF HOMOGENEOUS REGIONS IN SATELLITE IMAGE TIME SERIES

Pierre-Nicolas Mougel
pierre-nicolas.mougel@univ-nv.nc
PPME, Université de Nouvelle Calédonie
Avenue James Cook, Nouville 98851, New Caledonia

Nazha Selmaoui-Folcher
nazha.selmaoui@univ-nc.nc

ABSTRACT
Our work aims at analyzing satellites images time series using a pattern based data mining approach. We consider structures formed by a collection of regions sharing similar pixel properties in images. Such structure enable to discover hidden relations between disconnected regions corresponding to similar objects. The approach has been applied to the analysis of an area in New Caledonia and we present an example of pattern corresponding to several regions having an eroded soil.

1. INTRODUCTION AND STATE OF THE ART
During the last decade an increasing amount of high resolution satellite images sequences data became available. Their analysis made possible to monitor large areas and study their evolution over time. In particular, experts focused on regions with eroded soil as it has a major impact on the amount of cultivable land available. In order to assist experts in this task we propose a data mining approach to discover objects of interest in Satellites Images Time Series (SITS).

Several approaches have been proposed for the analysis of image time series. In [1, 2, 3] the authors propose methods based on physical models and consider the evolution of pixels. Other approaches applied to medical imagery focuses on object deformation, i.e., when the shape of objects change over time [4, 5]. In [6] several approaches are surveyed to discover scene transitions in movies. Regarding SITS analysis, pixel based approaches were proposed in [7, 8, 9, 10]. Their methods are based on sequential patterns mining to study meteorological phenomena. An object oriented approaches was proposed in [11] by first performing a segmentation of the images to discover objects and then building a dynamic graph representing objects evolution.

Previous data mining approaches for satellite image analysis consider either a single region (local pattern), or associate a class of object to each pixel in the image (classification). This last type of approach provides a good knowledge on the whole image but usually fail to discover structures at a local level. Regarding local patterns, we believe that considering not only a single region but several regions might give a better understanding of phenomena at a mesoscopic level. It might also help to discover non trivial relations between regions having similar behavior over time but not being necessarily geographically close. As such, our approach is similar to the one proposed in [12] where the authors consider collections of subgraphs.

Our pattern definition is based on the following assumptions: (1) an object consists in a set of connected pixels sharing similar properties, for example red, green, blue components or computed indexes, (2) several objects sharing the same properties might be found in an image, and (3) an object should exist in several consecutive timestamps. Given these assumptions, we propose a data mining approach to discover all the collections of objects sharing similar properties during at least two consecutive timestamps in a SITS. A formal definition of the pattern is given in Section 2.

To assist the experts during the analysis of a pattern, we propose to study the distribution of the properties in the surrounding area (spatial context) and previous / past timestamps (temporal context). As shown in Section 4, these distributions give a better understanding of the pattern evolution.

Our main contributions are as follow: In Section 2 we define a new class of pattern for the analysis of SITS. In Section 3 we present our enumeration strategy to compute the patterns. In Section 4 we present experimental results using satellite images of New Caledonia showing that our results uncover structures corresponding to existing objects, in particular regions with eroded soil. Section 5 briefly concludes.

2. CONTEXT AND DEFINITIONS
In this section, we formally define the context of our approach. We first propose a transformation of a SITS to a dynamic attributed graph then we define our patterns of interest.

---

This work was supported by the French National Research Agency through the project FOSTER (ANR-2010-COSI-012).

An object is usually defined as a region in an image corresponding to a real entity, e.g., a road, a lake.
2.1. Data setting formalization

We consider a SITS formed by images having the same dimension and where each pixel might be associated to a set of categorical attributes \( A = \{ A_1, \ldots, A_n \} \). From a SITS we build a dynamic attributed graph \( G = (V, E, A, AttVal) \), with \( V \) a set of vertices, \( E \) a set of undirected edges, and \( A \) a set of categorical attributes. Each vertex is associated to a pixel and \( V_{x,y} \) denotes the pixel at coordinate \((x, y)\). An edge is added between two vertices if their corresponding pixels are connected in the image including diagonal connectivity. Consequently, the graph has \((\text{image width} \times \text{image height})\) vertices. Function \( AttVal(V_{x,y}, A) \) associates to vertex \( V_{x,y} \) the value of attributes \( A \) at timestamp \( t \) if the attribute is present.

2.2. Pattern definition

In this setting, our objective is to discover several regions formed by vertices sharing the same attribute values at a given timestamp, and continue sharing similar attribute values (possibly different from previous ones) in at least one consecutive timestamp. First, we introduce the concept of region in a dynamic attributed graph.

**Definition 1 (Region).** A region is a set of vertices in a dynamic attributed graph such that there is a path between each pair of vertices.

Regions or vertices sharing the same attribute values will be considered as homogeneous. The following function returns the homogeneous attributes for a collection of regions.

**Definition 2.** The set of homogeneous attributes for all vertices in a collection of regions \( P = \{ R_1, \ldots, R_{|P|} \} \) at timestamp \( t \) is denoted \( HmgAtt(P) \), i.e., \( HmgAtt(P) = \{ A \in A \mid \exists a \in Dom(A) \text{ s.t. } \forall v \in \cup_{R \in P} AttVal(v, A), a = a \} \).

Our patterns of interest are named CoHRe which stands for Collection of Homogeneous Regions. They are formed by a collection of regions over several timestamps satisfying constraints on their attribute homogeneity over time and their size. They are defined as follows.

**Definition 3 (CoHRe).** Let \( G = (V, E, A, AttVal) \) be a dynamic attributed graph and \( h, g, \delta \) three user defined thresholds. A collection of regions \( P = \{ R_1, \ldots, R_{|P|} \} \) is a CoHRe if and only if:

- There is at least two consecutive timestamps in which the vertices of \( P \) share at least \( h \) attribute values, i.e., \( \exists t \text{ s.t. } |HmgAtt(P)_t| \geq h \wedge |HmgAtt(P)_{t+1}| \geq h \)
- The collection contains at least \( g \) regions having at least \( \delta \) vertices.

The first constraint enforces the homogeneity of the pattern w.r.t. the shared attributes. It also requires that this homogeneity remains in at least two consecutive timestamps. The second constraints allows to discard small patterns which might not be interesting for the experts. To set the thresholds, the analyst can select the lowest reasonable values and post-treat the collection without performing a new extraction to get the collection of CoHRe corresponding to any thresholds having higher values.

3. PATTERN ENUMERATION

In this section, we present the enumeration strategy used to compute the collection of CoHRe. Due to space limitation, we choose not to present the complete algorithm and to omit proofs. The enumeration strategy is based on the following property.

**Lemma 1.** Let \( P \) be a collection of regions. If \( P \) is a CoHRe then (1) there exist a sequence of collection of attribute values \( S = \{ a_0, \ldots, a_{|S|} \} \) starting at timestamp \( t \) and (2) \( P \) is the collection of connected components in \( G[V] \), where \( V \) is the set of vertices sharing all the attributes values in \( \cup_{0 \leq \Delta t < |S|} a_{\Delta t} \) at timestamp \( t + \Delta t \).

From Lemma 1 it is possible to compute all CoHRe by enumerating the sequences of attribute values along with the subgraphs induced by the set of vertices sharing all the attributes values. As the number of different sequences grows exponentially w.r.t. the number of attribute values and the number of timestamps, we propose several pruning techniques in order to reduce the search space.

According to the following property, we can avoid the enumeration of an attribute value shared by not enough vertices at a timestamp to contain a CoHRe.

**Lemma 2.** Let \( G \) be a dynamic attributed graph and \( x \) an attribute value shared by less than \( s \times g \) vertices in \( G \) at timestamp \( t \). Then, the graph induced by the vertices sharing attribute value \( x \) at timestamp \( t \) cannot contain a CoHRe.

The following property allows to reduce the subgraph under consideration. Along with Lemma 2, it can be used to reduce the search space.

**Lemma 3.** Let \( G \) be a graph. Only vertices in a connected components of \( G \) having at least \( s \) vertices can form a maximal CoHRe in \( G \) and all its subgraphs.

4. EXPERIMENTS

In this section we report experimental results. We first describe the dataset used then we illustrate the interest of the approach by means of a typical example of pattern corresponding to eroded soil.
4.1. Dataset

The dataset has been built using Landsat 7 satellite image of New Caledonia with the Landsat Enhanced Thematic Mapper Plus (ETM+) sensor. A Landsat 7 ETM+ satellite image consists of height bands, each band corresponding to the sensor response on different frequency ranges. We considered four bands, three bands with frequencies in the visible spectrum, the red, green and blue channels and one band corresponding to near infrared. On these bands the resolution is 30 meters per pixel. We considered a sequence of four images having dimension $355 \times 325$ (i.e., 115,375 pixels). The images were selected by keeping a time gap of at least 2 years and having a cloud cover under 5%. The corresponding SITS is presented on Figure 1 with a natural color rendering, i.e., only bands on the visible spectrum are used.

![Timestamp 0: December 2001](image1)

(a) Timestamp 0: December 2001  (b) Timestamp 1: March 2004

![Timestamp 2: April 2007](image2)

(c) Timestamp 2: April 2007  (d) Timestamp 3: September 2009

Fig. 1. Sequence of 4 satellite images of an area in the south of New Caledonia with date of acquisition.

In order to associate categorical attributes to the pixels we considered four numerical attributes: the Blue RGB Component (BC), the Normalized Difference Vegetation Index (NDVI) [13], the Redness Index (RI) [14] and the Brightness Index (BI) [15]. Given $r$, $g$, $b$, and $nir$ the response in respectively the red, green, blue and near infrared channels, the indexes are computed as $NDVI = \frac{nir - r}{nir + r}$, $RI = \frac{r^2}{g^3}$, and $BI = \sqrt{(r^2 + g^2 + b^2)/3}$. The NDVI is a green vegetation density indicator. The higher is the value, the more likely the area is supposed to contains dense green vegetation. The redness index indicates the presence of hematites in soil, one of several iron oxides. Finally, the brightness index is an indicator for naked soil.

Note that the red and green RGB components are not used in order to reduce attribute correlation as they are used to compute the indexes. The BC, NDVI, RI and BI numerical attributes are converted to categorical attributes using an equal frequency discretization. For each attribute, we built five classes labeled “−−”, “−”, “=”, “+”, “++”. The lowest attribute values are labeled “−−” up to the highest attribute values labeled “++”.

4.2. Example of CoHRe pattern

In this dataset, we considered CoHRe formed by at least $g = 6$ regions of at least $s = 100$ vertices sharing at least $h = 2$ common attribute values. With this thresholds we extracted 195 patterns. Figure 2 presents one of these patterns. It is formed by six regions homogeneous at timestamps 2 and 3. The homogeneous attributes are BI = “++”, RI = “++”, and NDVI = “−−” at timestamp 2 and BI = “++”, NDVI = “−−” at timestamp 3.

![Pattern at timestamp 2](image3)

(a) Pattern at timestamp 2  (b) Pattern at timestamp 3

Fig. 2. Example of CoHRe formed by six regions surrounded by red.

According to the experts, these regions corresponds to eroded soil. This is in good agreement with the homogeneous attribute values (i.e., naked soil and no green vegetation). The first row of Figure 3 presents the BI and NDVI attribute values distributions over all timestamps for the vertices in the pattern and the second row of Figure 3 presents the BI and NDVI attribute values distributions for the vertices in the neighborhood of the pattern. The pattern neighborhood is formed by the $k$-neighborhood of the vertices in the pattern without the vertices in the CoHRe with $k$ the minimal value such that the neighborhood is larger than the pattern. The distribution for attribute values BI = “++” and NDVI = “−−” are continually increasing since December 2001. This behavior seems to indicate that the erosion phenomenon is growing.

5. CONCLUSION

We proposed a data mining approach based on local patterns for the analysis of SITS. Our pattern definition allows to dis-

\[\text{It corresponds to a surface of } 6 \times 90,000 \text{ m}^2\]
cover collections of regions sharing similar properties in consecutive timestamps. The experimental results demonstrate that the approach is able to uncover structures corresponding to existing objects. In particular, we presented a pattern correspondence to several regions having eroded soil. The attribute values distribution seems to indicate that the erosion phenomena is increasing.

6. REFERENCES


Fig. 3. Attribute values distribution for the vertices in the pattern (first row) and in the neighborhood (second row).