EXTRACTION OF FREQUENT GROUPED SEQUENTIAL PATTERNS FROM SATELLITE IMAGE TIME SERIES

A. Julea[1,2,3]*, N. Méger*[1]*, C. Rigotti*[4]*, M-P. Doin*[5]*, C. Lasserre* [6]*, E. Trouvé*[1]*, P. Bolon*[1] *and V. Lăzărescu*[3]

[1]Université de Savoie/Polytech'Savoie, LISTIC Laboratory. B.P. 80439, F-74944 Annecy-le-Vieux Cedex, France
{andreea.julea|nicolas.meger|emmanuel.trouve|philippe.bolon}@univ-savoie.fr
[2]Institutul de Stiinte Spatiale. Bucharest, Romania
[3]Politehnica University of Bucharest/Faculty for Electronics, Telecommunications and Information Technology. Bucharest, Romania.
vl@elia.pub.ro
[4]Université de Lyon/INSA-Lyon, LIRIS Laboratory, CNRS - UMR 5205. 20, av. A. Einstein, F-69621 Villeurbanne Cedex, France
christophe.rigotti@insa-lyon.fr
[5] Ecole Normale Supérieure, Laboratoire de Géologie, CNRS - UMR 8538. 24 rue Lhomond, F-75231 Paris Cedex 05, France
doin@geologie.ens.fr
[6] Université Joseph Fourier, LGIT Laboratory, CNRS - UMR 5559. B.P. 53, F-38041 Grenoble Cedex 09, France
cecile.lasserre@ujf-grenoble.fr

ABSTRACT

This paper presents an original data mining approach for extracting pixel evolutions and sub-evolutions from Satellite Image Time Series. Those evolutions, namely the frequent grouped sequential patterns, are required to cover a minimum surface and to affect pixels that are sufficiently connected. These spatial constraints are actively used to face large data volumes and to select evolutions making sense for end-users. Successful experiments on an optical and a radar SITS are presented.

***Index Terms*—** satellite image times series, data mining, optical images, radar images, frequent grouped sequential patterns.

1. INTRODUCTION

Remote sensing techniques provide end-users with ever growing volumes of data. Indeed, the resolution of acquisitions is continually improved while the number of available channels also increases. In addition, acquisition rates have been boosted during the last few years. It is thus possible to gather large series of images concerning a given geographical zone. This kind of dataset is termed as a Satellite Image Time Series (SITS). The analysis of SITS raises new challenges as data volumes to be processed are huge and as both the temporal and the spatial dimensions have to be taken into account. Various techniques allowing to characterize evolutions in SITS have been proposed. Some of those techniques explore the data at the region level, more precisely, they extract regions from all the images so as to provide end-users with the evolutions of these regions (e.g., [1]). Other techniques link descriptors to each image of the SITS. A time sequence of descriptors is thus build and sub-evolutions that match temporal and frequency constraints are retained as the result (e.g., [2]). Pixel-based techniques have also been proposed, focusing either on specific evolution occurring at some time stamp, i.e., pixel change detection techniques (e.g., [3], [4]) or on the characterization of the whole sequence of pixel values and not of the sub-evolutions (e.g., synthetic channels-based techniques as proposed in [5] or clustering techniques as detailed in [6]). It is to notice that change detection techniques also work at the object/region level (but still needing assumptions about the type evolutions). Though similar to our approach, in the sense that generally both temporal and spatial dimensions are taken into account, none of these techniques can extract sets of grouped pixels sharing a same evolution or sub-evolution without first extracting objects/regions (e.g., [1, 2]) and/or without making any assumption about the type of evolution. For example, change detection techniques look for specific change classes while other pixel-based techniques only consider full evolutions and not sub-evolutions (e.g., [5, 6]). Furthermore, when searching for sub-evolutions, we aim at extracting them without giving any priority to any date of acquisition, which prevents us from using clustering tech-

The authors wish to thank the French Research Agency (ANR) for supporting this work through the EFIDIR project (ANR-2007-MCDC0-04, http://www.efidir.fr). They also thank the ADAM project and the CNES agency for making data available. Finally, the authors express their gratitude to Roxana Vintila (Research Institute for Soil Science and Agrochemistry - Bucharest, Romania) and to Gheorghe Petcu (National Agricultural Research and Development Institute Fundulea, Romania) for supplying the ground truth of the regions that we studied through acquisitions of the ADAM project.

niques. In [7], we presented a frequent sequential pattern-based approach that is preliminary to the one of this paper. It did not take into account the spatial grouping tendencies of the pixels that share a given evolution. It was thus possible to extract an evolution that holds for a lot of pixels that are not connected to each other. As a consequence, the evolutions provided to the end-user were sometimes difficult to interpret. Other works (e.g., [8, 9]) also rely on frequent sequential patterns to analyze spatio-temporal datasets. In [8], frequent sub-trajectories of objects, i.e. sequences of spatial locations sampled at consecutive timestamps, are mined. Trajectory mining can be performed only if trajectories are given as prior information, which requires objects to be identified. In [9], frequent sequential patterns are used to express spatio-temporal relations. Nevertheless, this requires end-users to set both temporal and spatial constraints. In this paper, no prior assumption about temporal information is made. This paper is organized as follows: Section2 introduces the concept of *frequent grouped sequential pattern* which is designed to extract meaningful pixel evolutions. It also details the manner in which such patterns can be efficiently extracted. Section 3 reports successful experiments on two different real datasets while Section 4 ends this paper.

2. FGS-PATTERNS

This paper presents an alternative and complementary approach, relying on evolution and sub-evolution extraction at the pixel level. A pixel evolution or sub-evolution is described using a sequential pattern, denoted $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_n$, where $A_1, A_2, \ldots, A_n$ are symbols representing discrete pixel states at n different dates which are not necessarily consecutive. Those patterns were initially proposed in [10] to mine sequences of commercial transactions. We intend to use those pixel evolutions and sub-evolutions, to find, in an unsupervised way, groups of pixels that could be of interest for end-users. In order to output pixel sets making sense both spatially and temporally, sets having at least σ pixels (i.e. a minimum surface) sharing the same temporal evolution α are selected. Pixels sharing α are said to *be covered* by α and are denoted $cov(\alpha)$. Furthermore, those same pixels are also required to exceed a minimum connectivity threshold κ. The connectivity measure that is used is called the *average connectivity*. It gives, for the pixels sharing α, the average number of neighbor pixels also sharing α. The 8 nearest neighbors (8-NN) are taken into account. Let us consider a *local connectivity function* $LC((x,y),\alpha)$ that returns, for a pixel (x,y), the number of neighbors covered by α. The average connectivity of α is then defined as follows: $AC(\alpha) = \frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{|cov(\alpha)|}$. Formally, an evolution (or sub-evolution) α is thus retained if $|cov(\alpha)| \geq \sigma$ and if $AC(\alpha) \geq \kappa$. In this case, it is called a *Frequent Grouped Sequential Pattern (FGS-pattern)*. FGS-patterns are a type of *frequent sequential patterns*. Indeed, a sequential pattern is frequent only if the constraint $|cov(\alpha)| \geq \sigma$ applies.

There are several algorithms for extracting frequent sequential patterns in a sound and complete way (e.g., [10, 11]). The main idea, used to reduce the execution times, is to take advantage of the *anti-monotonicity* property of the support. For an evolution/sub-evolution α, $support(\alpha) = |cov(\alpha)|$. This anti-monotonicity property can be informally stated as follows: if a sequential pattern α has a support λ then any pattern that contains at least the labels in α (also called *super-pattern*), in the same order, has a support equal to λ or lesser than λ. For example, if $support(D \rightarrow A) = \lambda$ then $support(D \rightarrow B \rightarrow A) \leq \lambda$. This property is commonly used by the sequential pattern extraction algorithms to limit the number of patterns to consider. For instance, if $D \rightarrow A$ has already been checked and found to be not frequent, then there is no need to test pattern $D \rightarrow B \rightarrow A$ since it cannot be frequent. Thanks to that property, a drastic reduction of the search space is made possible when looking for frequent sequential patterns.

The average connectivity constraint used to define FGS-patterns is not anti-monotonic, but it can be observed that for any frequent sequential pattern α since $|cov(\alpha)| \geq \sigma$, then $AC(\alpha) = \frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{|cov(\alpha)|} \leq \frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{\sigma}$ Thus a frequent pattern α that does not satisfy $\frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{\sigma} \geq \kappa$ cannot be a FGS-pattern. And, if we consider the conjunction of constraints $\mathcal{C} = support(\alpha) \geq \sigma \wedge \frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{\sigma} \geq \kappa$, this conjunction is anti-monotonic, since the value $\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)$ cannot increase for super-patterns of α. This conjunction can be thus actively used to prune the search space. We integrated the anti-monotonic conjunction $\mathcal{C}$ into the *PrefixGrowth* algorithm [11], that is a recent and efficient algorithm for sequential pattern mining under constraints. Beside checking $\mathcal{C}$ to prune the search space, the only required modification is to verify before outputting a pattern α that $AC(\alpha) \geq \kappa$, since satisfying $\mathcal{C}$ does not implies satisfying the average connectivity constraint. The implementation of the whole algorithm has been done in C using our own data structures.

3. EXPERIMENTS

All experiments have been run on a standard PC (Intel Core 2 @3GHz, 4 GB RAM, Linux kernel 2.6). Experiments on the ADAM (Data Assimilation by Agro-Modeling) [12] SITS are first reported. It is a SPOT SITS covering a rural zone in South Romania, near Bucharest which dedicated to the assessment of spatial data assimilation techniques within agronomic models. We selected 20 images between October 2000 and July 2001 containing 1000*1000 pixels each. They have been captured via three bands by SPOT satellites: B1 in green (0.5 - 0.59 μm), B2 in red (0.61 - 0.68 μm) and B3 in near in-

frared (NIR 0.78 - 0.89 μm). Their resolution is 20mx20m. For each pixel, and for each date, we consider a synthetic band B4. B4 is established by calculating the *Normalized Difference Vegetation Index* (NDVI) [13] using bands B2 and B3. B4 is thus defined as $B4 = \frac{B3-B2}{B3+B2}$. The NDVI index is widely used for detecting live green plant canopies in multispectral remote sensing data. An image quantization is performed by splitting the B4 value domain in 3 intervals that are equally populated. In order to minimize the influence of possible calibration defaults, quantization is separately done for each image. For a given acquisition date, a pixel is described by a single label that indicates which interval this pixel value belongs to. Label 1 relates to low NDVI values, label 2 represents mid NDVI values and label 3 denotes high NDVI values. Having at disposal the ground truth for the fields that belong to the Romanian National Agricultural Research and Development Institute (5.9% of the scene), we were able to evaluate our results. Parameter κ is to 6 to get pixels highly connected to each others and σ is set to 1% in order to ask for FGS-patterns relating to areas covering at least $4\ km^2$ (the whole image covers $400\ km^2$), i.e. relating to the main cultures of the scene. Extraction times do not exceed 600 seconds. In order to focus on the most specific evolution, maximal FGS-patterns are focused on, i.e. FGS-patterns in the output having no super-pattern also present in the output. We obtain 32 maximal patterns out of the 474 FGS-patterns that are extracted. One of these maximal patterns is $2 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$. The pixels covered by that pattern are represented in Figure 1b. According to the ground truth, it covers 61.4% of the pixels of the ground truth that relate to wheat culture, and 96.3% of the pixels it covers in the area where the ground truth is available, correspond to wheat culture. Interesting information can be drawn from such patterns. For instance, as it can be observed, some *holes* (small black areas) appear within the fields (large polygon almost completely filled in white) in Figure 1. The pixels of those holes are not covered by the pattern covering the ones in the white areas. Their temporal behavior is thus different from their surrounding pixels though they should be related to the same cultures. Some of those holes match pedological differences that have been reported by the experts while other holes are likely to be due to different fertilization and/or irrigation conditions. Such information is particularly interesting as it can be used to adapt locally soil fertilization or irrigation. Similar results are obtained for the other FGS-patterns. If occurrence dates are taken into account it is even possible to distinguish between the various species of a same type of culture. If shorter FGS-patterns, i.e. more general evolutions are considered, evolutions such as the one characterizing paths, fallows, cities and field borders or the one matching cultivated fields can be extracted.

The second dataset corresponds to interferograms and cover the lake Mead area, where the soil surface around the lake is affected by a subsidence/uplift motion that is correlated with water level fluctuations. We selected a subset of 20 interferograms obtained from images acquired between 1996 and 2008 and which have been computed using Synthetic Aperture Radar (SAR) images provided by the ERS satellites. Each interferogram gives the interferometric phase difference of its acquisition date relative to the master date 1995-10-08. The atmospheric phase screen of the master image is assigned to the master date. The analyzed images (759x716 pixels, 130m$\times$130m resolution) contain phase delays due to both atmospheric and deformation patterns. Phase delays were also quantized using 3 intervals. A strong positive value is interpreted as subsidence while a strong negative value relates to uplift. Setting σ to 10000 ($\sigma_{rel} \approx 2\%$) and κ to 6 provides 10173 FGS-patterns. In order to consider precise information, FGS-patterns having as many events as possible have been selected. We thus found 5 patterns having 15 events. The first labels of 4 patterns indicate important positive phase differences w.r.t. the master image and their last labels indicate important negative phase differences w.r.t. the same master image. Such patterns, called *water level-related patterns*, appear as correlated with water level fluctuations. The water level indeed increased between 1995-10-08 and 1998, while it decreased after 2000. In other words, those patterns suggest that there should be pixels for which subsidence (resp., uplift) is observed when the water level increases (resp., decreases). Such a behavior should be confirmed by a positive regression coefficient between phase delays and water level fluctuations. To check this assumption, we computed that regression coefficient using the whole interferometric dataset. Large positive regression coefficients are obtained on the localization of the water level-related patterns (see Figure 2). The regression coefficient is represented with a wrapped color scale (red/yellow/green/blue/violet). A positive (resp., negative) color cycle, from stable areas (i.e., image borders) to deformation zones, corresponds to a subsidence (resp., uplift) of 0.7 mm when the water level increases by 1 m. The remaining pattern relates to an evolution stating that

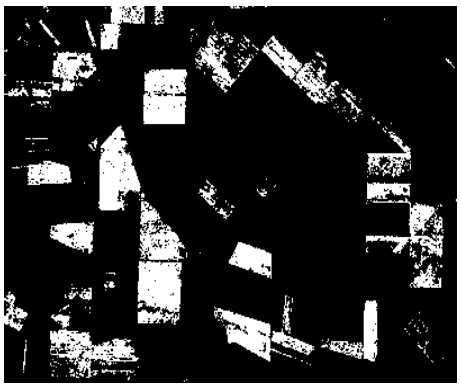

Fig. 1: Localization of pattern $2 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$.

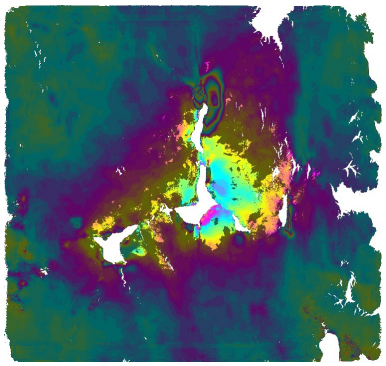

Fig. 2: Superposition of the joint localization of the water level-related patterns (enlightened) and regression coefficient (between phase delays and water level fluctuations).

pixels are affected by a continuous uplift. One of the largest area affected by this pattern is the Las Vegas one which is probably due to decreased water pumping in this part of Las Vegas aquifers. The localization of those five patterns well corresponds to zones where ground deformation is identified. Moreover, though atmospheric perturbations were present, none of these patterns report them, which demonstrates the ability of FGS-patterns to discard such random phenomena.

4. CONCLUSION

The original method presented in this paper is complementary to the existing techniques. In practice, it turns out to be effective in finding interesting groups of pixels, sharing meaningful common temporal evolutions, and that would not be exhibited by other approaches. The proposed approach is scalable and quite generic as successful experiments had been run on two real and large datasets: an optical and a radar SITS. Future work directions include using FGS-patterns to provide a single clustering of the whole SITS.

5. REFERENCES

[1] P. Héas and M. Datcu, "Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 7, pp. 1635– 1647, 2005.

[2] R. Honda and O. Konishi, "Temporal rule discovery for time-series satellite images and integration with RDB," in *PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, 2001, pp. 204–215, Springer-Verlag.

[3] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: a review," vol. 25, no. 9, pp. 1565–1596, May 2004.

[4] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Intl. J. of Remote Sensing*, vol. 25, no. 12, pp. 2365–2407, June 2004.

[5] E. Nezry, G. Genovese, G. Solaas, and S. Rémondière, "ERS - Based early estimation of crop areas in Europe during winter 1994-95," in *ERS Application, Proceedings of the Second International Workshop held 6-8 December 1995 in London*, Guyenne T.-D., Ed., 1996, vol. 383 of *ESA Special Publication*, p. 13.

[6] A. Ketterlin and P. Gançarski, "Sequence similarity and multi-date image segmentation," in *4th Intl Workshop on the Analysis of Multitemporal Remote Sensing Images*, Leuven, Belgique, july 2007.

[7] A. Julea, N. Meger, E. Trouve, and P. Bolon, "On extracting evolutions from satellite image time series," in *Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2008)*, Boston, MA, USA, 2008, vol. 5, pp. 228–231.

[8] Huiping Cao, Nikos Mamoulis, and David W. Cheung, "Mining frequent spatio-temporal sequential patterns," *Data Mining, IEEE International Conference on*, vol. 0, pp. 82–89, 2005.

[9] Yan Huang, Liqin Zhang, and Pusheng Zhang, "A framework for mining sequential patterns from spatio-temporal event data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 433–448, 2008.

[10] Rakesh Agrawal and Ramakrishnan Srikant, "Mining sequential patterns," in *Proc. of the 11th International Conference on Data Engineering (ICDE'95)*, Philip S. Yu and Arbee S. P. Chen, Eds., Taipei, Taiwan, 1995, pp. 3–14, IEEE Computer Society Press.

[11] Jian Pei, Jiawei Han, and Wei Wang, "Constraint-based sequential pattern mining: the pattern-growth methods," *Journal of Intelligent Information Systems*, vol. 28, no. 2, pp. 133–160, 2007.

[12] Centre National d'Etudes Spatiales, "Database for the Data Assimilation for Agro-Modeling (ADAM) project," online, http://kalideos.cnes.fr/index.php?id=accueil-adam.

[13] T.M. Lillesand and R.W. Kiefer, *Remote Sensing and Image Interpretation*, John Wiley and Sons, New York, fourth edition, 2000.