Discovering Spatio-Temporal Patterns in Precision Agriculture Based on Triclustering

Laura Melgar-García[1(✉)], Maria Teresa Godinho[2,3], Rita Espada[4], David Gutiérrez-Avilés[1], Isabel Sofia Brito[5,6], Francisco Martínez-Álvarez[1], Alicia Troncoso[1], and Cristina Rubio-Escudero[7]

[1] Data Science & Big Data Lab, Pablo de Olavide University, 41013 Seville, Spain
{lmelgar,dgutavi,fmaralv,atrolor}@upo.es
[2] Department of Mathematical and Physical Sciences, Polytechnic Institute of Beja, Beja, Portugal
mtgodinho@ipbeja.pt
[3] Center for Mathematics, Fundamental Applications and Operations Research, University of Lisboa, Lisbon, Portugal
[4] Associação dos Agricultores do Baixo Alentejo, Beja, Portugal
rita.espada.25@gmail.com
[5] Department of Engineering, Polytechnic Institute of Beja, Beja, Portugal
isabel.sofia@ipbeja.pt
[6] Instituto de Desenvolvimiento de Novas Tecnologias - Centre of Technology and Systems, Lisbon, Portugal
[7] Department of Computer Languages and Systems, University of Seville, Seville, Spain
crubioescudero@us.es

Abstract. Agriculture has undergone some very important changes over the last few decades. The emergence and evolution of precision agriculture has allowed to move from the uniform site management to the site-specific management, with both economic and environmental advantages. However, to be implemented effectively, site-specific management requires within-field spatial variability to be well-known and characterized. In this paper, an algorithm that delineates within-field management zones in a maize plantation is introduced. The algorithm, based on triclustering, mines clusters from temporal remote sensing data. Data from maize crops in Alentejo, Portugal, have been used to assess the suitability of applying triclustering to discover patterns over time, that may eventually help farmers to improve their harvests.

Keywords: Triclustering · Spatio-temporal patterns · Precision agriculture · Remote sensing

1 Introduction

It is a well-established fact that shortage of natural resources endangers our future. Public awareness of these problems urges local authorities to intervene and impose tight regulations on human activity. In this environment, reconciling economic and environmental objectives in our society it is mandatory.

Precision agriculture (PA) has an important role in the pursuit of such aspiration, as the techniques used in PA permit to adjust resource application to the needs of soil and crop as they vary in the field. In this way, specific-site management (that is the management of agricultural crops at a spatial scale smaller than the whole field) is a tool to control and reduce the amount of fertilizers, phytopharmaceuticals and water used on site, with both ecological and economic advantages. Indeed, being able to characterize how crops behave over time, extracting patterns and predicting changes is a requirement of utmost importance for understanding agro-ecosystems dynamics [1].

One of the major concerns associated to the shortage of natural resources is the enormous consumption of water associated to farming activities. Water is a scarce resource worldwide and this problem is particularly acute in the South of Europe, where the Alentejo (Portugal) and Andalusia (Spain) regions are located. Both regions are mainly agriculture-dependent and thus farmers and local authorities are apprehensive about the future.

In this paper, an algorithm is proposed to delineate management zones by measuring the variability of crop conditions within the field through the analysis of time series of geo-referenced vegetation indices, obtained from satellite imagery. In particular, the well-known normalized difference vegetation index (NDVI), indicator for vegetation health and biomass, is used to analyze how the crop varies over time in order to find patterns that may help to improve its production. There are more vegetation indices as GNDVI, SAVI, EVI or EVI2 [2,3] which should be used in extended works.

A triclustering method, based on an evolutionary strategy called TriGen [4] has been applied to a set of satellite images indexed over time from a particular maize crop in Alentejo, Portugal. Although the method was originally designed to discover gene behaviors over time [5], it has also been applied to other research fields such as seismology [6]. The TriGen is a genetic algorithm, and therefore the fitness function is a key aspect since it leads to the discovery of triclusters of different shapes and aspects. The multi-slope measure (MSL) [7], the three-dimensional mean square residue (MSR3D) [8] and the least squared lines (LSL) [9] are the available fitness functions to mine triclusters in TriGen. Furthermore, the TRIclustering quality (TRIQ) index [10] was proposed to validate the results obtained from the aforementioned fitness functions.

The rest of the paper is structured as follows. In Sect. 2, the recent and related works are reviewed and the process of data acquisition and preprocessing is described. In Sect. 3 the proposed algorithm and its adaption to this particular problem are described. In Sect. 4 the results are presented and discussed. Finally, in Sect. 5, the conclusions of this work and point directions for future work are presented.

2 Related Works

This section reviews the most recent and relevant works published in the field of spatio-temporal patterns in precision agriculture.

The spatio-temporal pattern discovery issues for satellite time series images are discussed in [11]. The authors introduced how to perform an automatic analysis of these patterns and the problem of determining its optimal number. Unfortunately, these questions are still open issues in the literature and it is unlikely that a general consensus can be reached in the near future.

The estimation of spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series was analyzed in [12]. Four different approaches were applied to eight years of Australian crops, including calculation of temporal mean and standard deviation layers, spatio-temporal key NDVI patterns, different climatic variables and relationships between productivity and production.

In Fung et al. [13], the authors proposed a novel spatio-temporal data fusion model for satellite images using Hopfield Neural Networks. Synthetic and real datasets from both Hong Kong and Australia, respectively, were used to assess the method performance, showing remarkable results and outperforming some of other existing methods.

The use of convolutional neural networks (CNN) is being currently applied in a wide range of spatio-temporal patterns discovery applications [14]. Hence, Tan et al. [15] enhanced an existing CNN model for image fusion by proposing a new network architecture and a novel loss function. Results showed superior performance in terms of accuracy and robustness. Ji et al. [16] proposed a 3D CNN dealing with multi-temporal satellite images. In this case, the method was designed for crop classification. After discussing the results achieved, outperforming existing well-established methods, the authors claimed that it is especially suitable for characterizing crop growth dynamics.

An ensemble model for making spatial predictions of tropical forest fire susceptibility using multi-source geospatial data can be found in [17]. The authors evaluated the Lao Cai region, Vietnam, through several indices including NDVI.

Bui et al. [18] proposed an approach based on deep learning for predicting flash flood susceptibility. Real data from a high frequency tropical storm area were used to assess its performance.

Clustering-based approaches with application to precision agriculture can also be found in the literature. Thus, clustering tools for integration of satellite imagery and proximal soil sensing data are described in [19]. In particular, a novel method was introduced with the aim of determining areas with homogeneous parts in agricultural fields.

The application of triclustering to georeferenced satellite images time series can be also found in [20]. However, the authors addressed a different problem: the patterns analysis of intra-annual variability in temperature, using daily average temperature retrieved from Dutch stations spread over the country.

3 Methodology

This section introduces the *TriGen* algorithm, the methodology used to extract behavior patterns from satellite images along with the time points when they were taken. This methodology is applied to a 3D dataset (composed of rows, columns, and depths) that represents the X-axis coordinates (rows) and the Y-axis coordinates (columns) of each satellite image taken at a particular instant (depth). *TriGen* is a genetic algorithm that minimizes a fitness function to mine subsets of X-axis coordinates, Y-axis coordinates, and time points, called triclusters, from 3D input datasets. The *NDVI* values in the yielded subsets of $[X, Y]$ coordinates along with the subset of time points, share similar behavior patterns.

In general terms, *TriGen* is explained from two main concepts, presented in the following sections: the triclustering model applied to the case study (Sect. 3.1) and the inputs, output and algorithm workflow of *TriGen* (Sect. 3.2).

3.1 Triclustering

The case study presented has been modeled as a triclustering problem, in which 3-dimensional patterns are extracted from an original dataset. Prior to explaining this development, it is necessary to distinguish between two types of dataset:

- D_{2D} (2-dimensional dataset): a matrix with a set of instances (rows) and a set of features (columns).
- D_{3D} (3-dimensional dataset): a 3D matrix with a set of instances (rows) and features (columns), taken at a particular time points (depths).

Clustering algorithms are applied to D_{2D} datasets performing a complete partition it; for each yielded clusters, the values of the grouped instances share a behavior pattern through all features. In contrast, the triclustering algorithms work with D_{3D} datasets and group not only subsets of instances, but also subsets of features and time points. In this case, for each yielded tricluster, the values of grouped instances for the particular grouped features share a behavior pattern through a group of time points.

Thus, for this case study, the application of the *TriGen* algorithm to a D_{3D} dataset of satellite images where the instances are the Y coordinates of the space, the features are the X coordinates of the area and, the time points are the moment at the images where taken, will yield a set of triclusters representing, each of them, a behavior pattern of NDVI, for a particular subspace (subset of Y and X coordinates) through a specific set of times (subset of time points).

3.2 The *TriGen* Algorithm

In order to mine the triclusters from the D_{3D} dataset of satellite images, the *TriGen* algorithm is applied. *TriGen* is based on the genetic algorithm paradigm;

therefore, it evolves a population of individuals employing genetic operators during a specific number of generations to optimize an evaluation function.

The inputs of *TriGen* are two: the D_{3D} dataset of satellite images and the initial configuration of the genetic process. The parameters that can be set are the number of triclusters to mine (N), the number of generations of the genetic process (G), the size of the initial population (I), the fraction of population that promoted to the next generation (Sel) and, the probability of mutation (Mut). A complete analysis of the influence of these parameters in the performance of the algorithm can be consulted in [4,7,8].

Each individual in the genetic process is represented as a tricluster and composed of a subset of instances of D_{3D}, a subset of features of D_{3D} and, a subset of time points of D_{3D}; the individuals (triclusters) with the best fitness function value are the output of the algorithm.

The genetic operators allow for searching among the individuals to obtain better solutions for each generation. For the *TriGen* algorithm, the description of them is the following:

- Initial population. The individuals are generated with three methods. The first method consists in a random selection of the elements of the individuals. The second one, considering the rows and columns of D_{3D} as a geographical area, performs a random selection of a rectangular sub-area and time points. The last one selects the elements of the individuals taking into account the rows, columns, and time points of D_{3D} visited in already extracted solutions in order to explore the most number of elements of D_{3D}.
- Evaluation. This operator applies the fitness function to the population in order to asses the quality of each individual. The fitness function used in the present case study is *MSL*.
- Selection. A tournament selection algorithm is applied to promote the individuals with the best evaluation to the next generation. The rest of individuals in the next population are generated by crossing and mutations.
- Crossover. Two individuals are combined to generate another two ones. The crossover used is the one point crossing. Each of the three elements of the two involved individuals (parents), are split in two and the four parts are combined two new individuals (offspring).
- Mutation. This operator modifies an individual to obtain variability in the next generation. Three actions have been used: insertion of a new coordinate $[X, Y]$ or time point, deletion of an existing coordinate $[X, Y]$ or time point and change of an existing coordinate $[X, Y]$ or time point.

4 Results

This section reports and discusses the results achieved after the application of the proposed methodology to a particular dataset. Thus, Sect. 4.1 describes the high resolution remote sensing imagery used in this study and Sect. 4.2 introduces the validation function used to evaluate the quality of the triclusters obtained. Finally, Sect. 4.3 reports the spatio-temporal patterns obtained and discusses its physical meaning.

4.1 Dataset Description

Located in the Baixo Alentejo region of Portugal, the site under study is a 63.82 ha maize plantation, with center at coordinates $(38°08'12''N, 7°53'42''W)$, as shown in Fig. 1. The site was monitored between sowing (April of 2018) and harvesting (September of the same year) and it is characterized by a set of nineteen images retrieved at time intervals of five, ten and fifteen days, from the Sentinel 2 Mission. The research site was irrigated using a central pivot irrigation system.

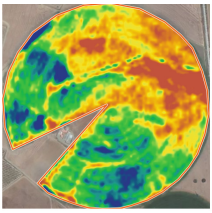
Fig. 1. Location of the research site.

Vegetation indices are, by definition, algebraic combinations of the measured canopy reflectance of different wavelength bands [21]. The use of Vegetation Indices in this context is based on the fact that healthy and unhealthy plants reflect light differently. Due to this difference, crop canopy multispectral reflectance, which is detectable remotely through aerial or satellite imagery, can be used to monitor the state of the crop [22]. For these reasons, one of the most widely used indices is applied to the images: the Normalized Differential Vegetation Index (NDVI). The NDVI can be calculated as follows:

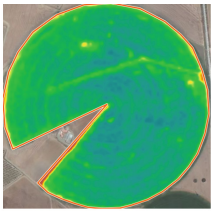
$$NDVI = \frac{NIR - Red}{NIR + Red}, \tag{1}$$

where Red and NIR stand for the spectral reflectance measurements acquired in the red (visible) and near-infrared regions, respectively, and $NDVI \in [-1, 1]$.

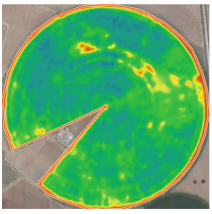
As pointed out in [23], the NDVI index has proven to be quite useful in monitoring variables such as crop nutrient deficiency, final yield in small grains, and long-term water stress. All these variables are very important to the case study presented here. Figure 2 illustrates how the NDVI of the target area varies over time, including images at six different chronologically ordered time stamps.



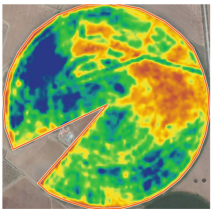
(a) June 19^{th}, 2018



(b) August 3^{rd}, 2018



(c) September 22^{nd}, 2018



(d) October 17^{th}, 2018

Fig. 2. Sample NDVI values for the research site, chronologically ordered.

4.2 Behaviour Patterns Quality, the *TRIQ* Measure

The *TRIQ* index has been used in order to measure the quality of the yielded triclusters in this case study, that is, the quality of the behavior pattern that a tricluster depicts. *TRIQ* measures the quality of a tricluster based on three elements: the similarity of the behavior patterns of the grouped $[X, Y]$ points along with the grouped time points and the Pearson's and Spearman's correlation indexes between all the $[X, Y]$ time series of the tricluster. $TRIQ$ values rank in the $[0, 1]$ interval; $TRIQ$ is a measure to maximize. A full description, definition, development, and performance of *TRIQ* can be consulted in [10].

4.3 Discovery of Spatio-Temporal Patterns in Maize Crops

TriGen analyzes the evolution of NDVI indices in each specific area and discovers triclusters of similar behavior patterns. Thus, the dataset with the NDVI indices of the satellite images over time is the first input of the algorithm.

TriGen has some configuration parameters, above-mentioned in Sect. 3.2. The algorithm has been run several times with different settings for each parameter. The configuration parameters that fit the best to these images are: $G = 10$, $I = 200$, $Sel = 0.8$ and $Mut = 0.1$. The number of triclusters to find is 4 and the fitness function used is MSL. Therefore, these values are the second input of the algorithm.

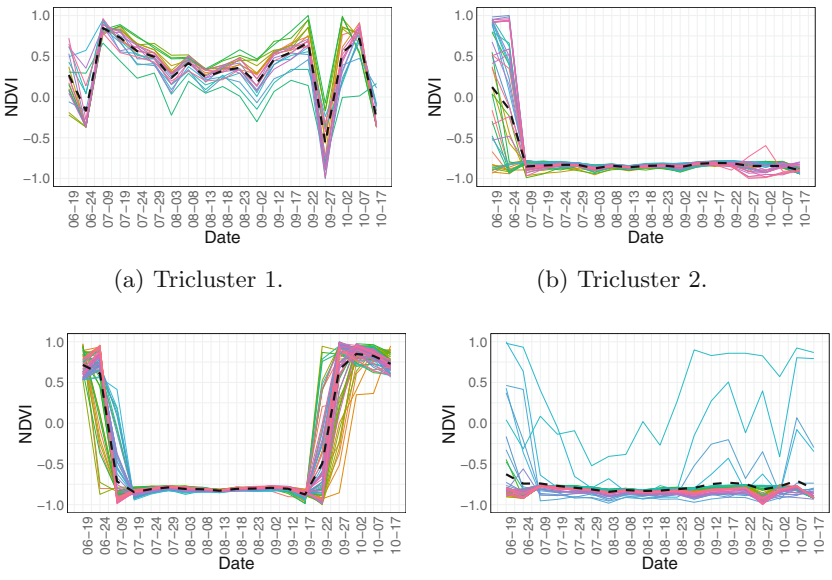
Each of the 4 discovered triclusters has a $TRIQ$ measure. The first one has a $TRIQ$ of 0.803, the second has 0.753, the third has 0.827 and the fourth has 0.742. These high values lead to confirm the good quality of all the triclusters. However, this measure itself does not guarantee the meaningfulness of the triclusters discovered. In order to interpret the evolution of the triclusters in an

accurate way, field's farmers provided additional information about the plantations site-specific conditions, such as irrigation or fungicide, for the same period. This information confirmed that triclusters were meaningful also in geophysical terms.

The triclusters discovered are represented in Figs. 3a, 3b, 3c and 3d. Each graph represents the evolution of the NDVI of the selected $[X, Y]$ components over time. The black dashed line added in each graph represents the mean value of all components. Triclusters components share a similar behavior. The first tricluster corresponds to areas with high NDVI values that remain almost constant over time. The components of the second tricluster are fields that start with a high NDVI and experiment a sudden decrease for the rest of the dates studied. The beginning of the third tricluster is similar to the previous one but with a recovery of the initial values after mid September. The last tricluster is formed by areas with constant low NDVI over time.

The changes of the NDVI values identified by triclusters 1, 2 and 3 during the first samples seem to be related with the use of fertilizers and the increase of the amount of water for the irrigation process. The third tricluster and some components of the first one show a change in their behaviour at mid September. It could be related to the application of fungicide by the farmers during August.

The proposed algorithm contributes in finding areas of similar crop conditions over the NDVI vegetation index using satellite images in different times. In addition, as TriGen includes the time dimension, the evolution over time of



(a) Tricluster 1.

(b) Tricluster 2.

(c) Tricluster 3.

(d) Tricluster 4.

Fig. 3. Triclusters found by *TriGen* in 2018.

each tricluster's features can be analyzed. Nevertheless, the interpretation of the results needs the validation of a specialist as the TRIQ measure does not consider neither geographical nor environmental features.

5 Conclusions

The suitability of applying triclutstering methods to discover spatio-temporal patterns in precision agriculture has been explored in this work. In particular, a set of satellite images from maize crops in Alentejo, Portugal, has been analyzed in terms of its NVDI temporal evolution. Several patterns have been found, identifying zones with tendency to obtain greater production and others in which human interventions are required to improve the soil properties. Several issues remain unsolved and are suggested to be addressed in future works. First, these patterns may help to identify the most suitable moments to apply fertilizers or pesticides. Second, the forecasting of maize production could be done based on such patterns. Third, additional crop production features such as amounts and characteristics of the fertilizers, phytopharmaceuticals and water used throughout the season (moister probes placed 30 cm underground were used to access the soil need for water before irrigation, when needed), would help to discover more robust patterns. Fourth, more images records during more years and a specific measure to assess the quality and meaning of precision agriculture triclusters would improve the application of the proposed algorithm to agricultural production. Fifth, more vegetation indices should be used.

Acknowledgements. The authors would like to thank the Spanish Ministry of Economy and Competitiveness for the support under project TIN2017-88209 and Fundação para a Ciência e a Tecnologia (FCT), under the project UIDB/04561/2020. The authors would also like to thank António Vieira Lima for giving access to data and Francisco Palma for his support to the whole project.

References

1. Tan, J., Yang, P., Liu, Z., Wu, W., Zhang, L., Li, Z., You, L., Tang, H., Li, Z.: Spatio-temporal dynamics of maize cropping system in Northeast China between 1980 and 2010 by using spatial production allocation model. J. Geog. Sci. **24**(3), 397–410 (2014)
2. Jurecka, F., Lukas, V., Hlavinka, P., Semeradova, D., Zalud, Z., Trnka, M.: Estimating crop yields at the field level using landsat and modis products. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis **66**, 1141–1150 (2018)
3. Jiang, Z., Huete, A., Didan, K., Miura, T.: Development of a two-band enhanced vegetation index without a blue band. Remote Sens. Environ. **112**, 3833–3845 (2008)
4. Gutiérrez-Avés, D., Rubio-Escudero, C., Martínez-Álvarez, F., Riquelme, J.C.: Trigen: A genetic algorithm to mine triclusters in temporal gene expression data. Neurocomputing **132**, 42–53 (2014)

5. Melgar, L., Gutiérrez-Avilés, D., Rubio-Escudero, C., Troncoso, A.: High-content screening images streaming analysis using the STriGen methodology. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 537–539 (2020)
6. Martínez-Álvarez, F., Gutiérrez-Avilés, D., Morales-Esteban, A., Reyes, J., Amaro-Mellado, J.L., Rubio-Escudero, C.: A novel method for seismogenic zoning based on triclustering: application to the Iberian peninsula. Entropy **17**(7), 5000–5021 (2015)
7. Gutiérrez-Avilés, D., Rubio-Escudero, C.: MSL: a measure to evaluate three-dimensional patterns in gene expression data. Evol. Bioinform. **11**, 121–135 (2015)
8. Gutiérrez-Avilés, D., Rubio-Escudero, C.: Mining 3D patterns from gene expression temporal data: a new tricluster evaluation measure. Sci. World J. **2014**, 1–16 (2014)
9. Gutiérrez-Avilés, D., Rubio-Escudero, C.: LSL: a new measure to evaluate triclusters. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, pp. 30–37 (2014)
10. Gutiérrez-Avilés, D., Giráldez, R., Gil-Cumbreras, F.J., Rubio-Escudero, C.: TRIQ: a new method to evaluate triclusters. BioData Min. **11**(1), 15 (2018)
11. Radoi, A., Datcu, M.: Spatio-temporal characterization in satellite image time series. In: Proceedings of the International Workshop on the Analysis of Multitemporal Remote Sensing Images, pp. 1–4 (2015)
12. Hill, M.J., Donald, G.E.: Estimating spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series. Remote Sens. Environ. **84**(3), 367–384 (2003)
13. Fung, C.H., Wong, M.S., Chan, P.W.: Spatio-temporal data fusion for satellite images using Hopfield neural network. Remote Sens. **11**(18), 2077 (2019)
14. Kamilaris, A., Prenafeta-Boldú, F.: A review of the use of convolutional neural networks in agriculture. J. Agric. Sci. **156**(3), 312–322 (2018)
15. Tan, Z., Di, L., Zhang, M., Guo, L., Gao, M.: An enhanced deep convolutional model for spatiotemporal image fusion. Remote Sens. **11**(18), 2898 (2019)
16. Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y.: 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. Remote Sens. **10**(1), 75 (2018)
17. Tehrany, M.S., Jones, S., Shabani, F., Martínez-Álvarez, F., Bui, D.T.: A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data. Theoret. Appl. Climatol. **137**, 637–653 (2019)
18. Bui, D.T., Hoang, N.-D., Martínez-Ávarez, F., Ngo, P.-T.T., Hoa, P.V., Pham, T.D., Samui, P., Costache, R.: A novel deep learning neural network approach for predicting flash flood susceptibility: a case study at a high frequency tropical storm area. Sci. Total Environ. **701**, 134413 (2020)
19. Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., Aspinall, D., Ji, W.: Clustering tools for integration of satellite remote sensing imagery and proximal soil sensing data. Remote Sens. **11**(9), 1036 (2019)
20. Wu, X., Zurita-Milla, R., Izquierdo-Verdiguier, E., Kraak, M.-J.: Triclustering georeferenced time series for analyzing patterns of intra-annual variability in temperature. Ann. Am. Assoc. Geogr. **108**, 71–87 (2018)

21. Schueller, J.: A review and integrating analysis of spatially-variable control of crop production. Fertil. Res. **33**, 1–34 (1992)
22. Xue, J., Su, B.: Significant remote sensing vegetation indices: a review of developments and applications. J. Sens. **17**, 1353691 (2017)
23. Govaerts, B., Verhulst, N.: The normalized difference vegetation index (NDVI) GreenSeekerTM handheld sensor: toward the integrated evaluation of crop management. CIMMYT (2010)