A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture

Laura Melgar-García [a], David Gutiérrez-Avilés [b], Maria Teresa Godinho [c,d], Rita Espada [e], Isabel Sofia Brito [f,g], Francisco Martínez-Álvarez [a,*], Alicia Troncoso [a], Cristina Rubio-Escudero [b]

[a] Data Science & Big Data Lab, Pablo de Olavide University, ES-41013 Seville, Spain
[b] Department of Computer Science, University of Seville, Avda. Reina Mercedes s/n, Seville 41012, Spain
[c] Department of Mathematical and Physical Sciences, Polytechnic Institute of Beja, Portugal
[d] Center for Mathematics, Fundamental Applications and Operations Research, University of Lisboa, Portugal
[e] Associação dos Agricultores do Baixo Alentejo, Beja, Portugal
[f] Department of Engineering, Polytechnic Institute of Beja, Portugal
[g] Instituto de Desenvolvimiento de Novas Tecnologias – Centre of Technology and Systems, Lisboa, Portugal

Keywords:

Big data triclustering

Precision agriculture

Spatio-temporal patterns

A B S T R A C T

Precision agriculture focuses on the development of site-specific harvest considering the variability of each crop area. Vegetation indices allow the study and delineation of different characteristics of each field zone, generally invisible to the naked-eye. This paper introduces a new big data triclustering approach based on evolutionary algorithms. The algorithm shows its capability to discover three-dimensional pat-terns on the basis of vegetation indices from vine crops. Different vegetation indices have been tested to find different patterns in the crops. The results reported using a vineyard crop located in Portugal depicts four areas with different moisture stress particularities that can lead to changes in the management of the vineyard. Furthermore, scalability studies have been performed, showing that the proposed algorithm is suitable for dealing with big datasets.

1. Introduction

It is a well-established fact that the era of Big Data [1] has changed the way in which data are generated, stored and processed, to the extent that 90% of the data that exist in the world has been generated during the last years [2]. These vast amount of data can be difficult to understand or even to analyze, and therefore the need for techniques to process this information arises. In this sense, new tools have been developed under the title of Data Science [3].

One of the areas that benefits from these developments is Precision Agriculture (PA), that can be defined as the application of technologies and principles to manage spatial and temporal variability associated to all aspects of agricultural production for the purpose of improving crop performance and environmental quality [4]. It is a 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. 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 [5].

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, a new algorithm, hereinafter called bigTriGen, is proposed to delineate management zones by measuring the variability of crop conditions within the field. For this purpose, bigTriGen analyzes time series of geo-referenced vegetation indices,

* Corresponding author.
E-mail address: fmaralv@upo.es (F. Martínez-Álvarez).

obtained from satellite imagery. Thus, the bigTriGen algorithm, based on the evolutionary strategy introduced in the TriGen algorithm [6], is a triclustering method capable to analyze a set of satellite images indexed over time in addition to the ability to analyze vast three-dimensional datasets in a big data environment. It has been applied to a vineyard crop located in Baixo Alentejo, Portugal, with different experimental datasets in order to test its scalability.

The rest of the paper is structured as follows. In Section 2, the recent and related works are reviewed. In Section 3 our proposal is described. In Section 4 the results obtained using the vineyard crop dataset are presented and discussed. Finally, in Section 5, the conclusions of this work and point directions for future works are presented.

2. Related works

Interest in precision agriculture methods applied to viticulture has had a tremendous growth in the last decade: at the research level, the number of papers published has increased from 20 in 2011 [7] to 517 hits in response to googling "vineyard precision agriculture" in Google Scholar, in spite of having restricted the search to the current year. In fact, generally, vineyards meet the three classical conditions that are required in order to site specific management methods to be justified: (1) significant spatial variability within field exists (2) the causes of this variability can be identified and measured, and (3) the information from these measurements can be used to modify crop-management practices to increase profit and quality and decrease environmental impact [8]. These three conditions define themselves three important lines of research that complement each other. This paper addresses the first one, that is, we aim at identifying areas within the field with different behaviors as to grape quality and productivity. This objective involves both gathering data and extracting information from data. In the following, some of the proposed methods to deal with these topics are reviewed.

Vineyards are often planted in irregular/steep terrains resulting in difficult and expensive direct inspection tasks for wine growers. Thus, discrete point sampling, which is the most traditional mean of data collection on soil conditions and/or plant growth and development, is very difficult to implement on this type of crop [8,9]. On the other hand, aerial remote sensors have proven to be very effective means of collecting data as they can provide, at a relatively low cost, a fairly detailed, spatially referenced measure of almost all the same features. Both satellite and airborne imaging systems, namely unmanned aerial vehicles (UAV), with multispectral and hyperspectral cameras have been used for gathering crop related data for a few decades. Several papers have reviewed the use of aerial remote sensors and compared the quality of the information extracted from both means in accessing vineyard variability [10–13]. Satellite imagery is affordable and easily available, but inferior with regard to resolution and more vulnerable to atmospheric interference. Nevertheless, although it is widely accepted that UAV imagery provides a more complete view of the field [10–12], it has also been shown that good correlation exists between Normalized Difference Vegetation Index (NDVI) data from Sentinel 2 and NDVI unfiltered data from UAV [13]. Additionally, [14] shows that by complementing aerial data with information gathered by ground-based sensors, high-resolution management zones can be delineated.

Rather comprehensive reviews on methods for data analysis in precision agriculture are presented in [15,16]. The identification of site-specific management zones is achieved mostly through clustering techniques. Twenty of those techniques are compared in [17]. The comparison was conducted with data obtained between 2010 and 2015 from three commercial agricultural fields cultivated with soya bean and maize in Brazil. Then, the divisions suggested by the results of a one-way ANOVA performed on the yields were compared to the divisions obtained using the various algorithms. The results showed that 17 out of the 20 produced quite good results, although McQuitty's Method and Fanny were considered to be the best choices. [18] presents a smaller study on four unsupervised methods applied to vineyard canopy segmentation in three different scenarios, with both RGB (Red-Green–Blue) and NRG (Near Infrared-Red-Green) imagery. The k-means algorithm has proven to be the more stable over the identification in the orthomosaic and sub-regions regarding the RGB acquisitions, whereas the HSV-RGN algorithm is the more stable over the identification in the orthomosaic and sub-regions regarding the NRG acquisitions. Many other studies are available, where a given method is proposed to define management zones in vineyards, based in various characteristics of the crop (disease detection, berry composition and sanitary status under humid conditions, among others) but we are not aware of the existence of a wider recent comparison on that matter.

Clustering tools to determinate time space patterns in precision agriculture can also be found in the literature: the evapotranspiration of a Pinot noir commercial vineyard in California was characterize through the unsupervised fuzzy c-means algorithm in [19] and, in [20], NDVI spatio-temporal patterns were obtained for a corn field in the Alentejo, Portugal, by means of a triclustering methodology.

Triclustering methodology has become a very researched area in the last years [21]. Some algorithms are based on genetic operators as [6] that included different evaluation measures [22–25] or [26] which used COVID-19 propagation model to optimize multi-objective functions. The characteristic of mining spatio-temporal patterns can be applied to different study areas as: medical [27], seismic [28] or even in environmental sensors in online learning [29]. Regarding the big data characteristic, [30] introduced a new parallel batch algorithm based on k-means providing speed results. [31] presented a parallel and scalable validation model for simple clusters in big data using Apache Spark. However, there is still much research to conduct in the development of big data triclustering algorithms.

3. Methodology

In this section, the methodology in order to obtain triclusters that enclose patterns from crops images is presented. Firstly, the triclustering that models the problem is described in Section 3.1, and finally, the way in which triclustering is applied, that is, the bigTriGen algorithm is presented in Section 3.2.

3.1. Problem modeling: triclustering

The triclustering techniques emerge as an evolution of clustering techniques applied over three-dimensional ($3D$) datasets. Triclustering aims at obtaining a set of triclusters ($3D$ clusters) from the input dataset, with the values of each tricluster representing a pattern of behavior.

To formalize the triclustering concepts, firstly a three-dimensional dataset D_{3D} composed of the three sets D^I, D^F and D^{TP} is defined as follows:

$$D_{3D} = \left\{D^I, D^F, D^{TP}\right\} \tag{1}$$

where $D^I = \{i_1, i_2, \ldots, i_I\}$ represents the I instances, $D^F = \{f_1, f_2, \ldots, f_F\}$ the F features and $D^{TP} = \{t_1, t_2, \ldots, t_{TP}\}$ the TP time points of the dataset.

Each pair instance-feature of the dataset represents a time series D_{TS}, that is, a sequence of time-indexed values from the time instant t_1 to t_{TP} as shown in the following equation:

$$D_{TS}(i,f) = \{v_{t_1}, v_{t_2}, \dots v_{t_{TP}}\}, \quad \forall i \in D^I, \quad \forall f \in D^F, \quad \forall t \in D^{TP} \tag{2}$$

In conclusion, D_{3D} is typically arranged as a data cube where the rows are the instances, the columns are the features and the depths are the time points of the time series.

Secondly, a tricluster T is a subset of instances T^I, features T^F and time points T^{TP} of D_{3D} defined by the following equation:

$$T = \left\{T^I, T^F, T^{TP}\right\} \text{with} T^I \subset D^I \quad T^F \subset D^F \quad T^{TP} \subset D^{TP} \tag{3}$$

The time points in T^{TP} for a particular instance and feature of the tricluster make up a continuous and ordered sub-sequence of values of the entire sequence of the dataset, that is, a time series T_{TS} from initial time t_s (first tricluster time point) to final time t_{TS} (last tricluster time point) defined as follows:

$$T_{TS}(i,f) = \{v_{t_s}, v_{t_{s+1}}, \dots v_{t_{TS}}\}, \quad \forall i \in T^I, \quad \forall f \in T^F, \quad \forall t \in T^{TP} \tag{4}$$

Thus, the behavior patterns (BP) depicted by each time series of the tricluster will present similar behavior regarding the values or tendency. To summarize, a tricluster is a subset of instances, features, and time points of a three-dimensional dataset,with time series that depict a similar behavior pattern.

$$BP(T_{TS}(i_A,f_A)) \sim BP(T_{TS}(i_B,f_B)), \quad \forall i_A, i_B \in T^I, \quad \forall f_A, f_B \in T^F \tag{5}$$

Finally, a triclustering model of a three-dimensional dataset, $M_{D_{3D}}$, is a set of N triclusters defined as:

$$M_{D_{3D}} = \{T_1, T_2, \dots, T_N\} \tag{6}$$

3.2. The bigTriGen algorithm

bigTriGen is applied to obtain a triclustering model providing a set of behavior patterns from the input dataset. bigTriGen is based on the paradigm of genetic algorithms. In that sense, a complete evolutionary process is performed for each tricluster to be obtained, i.e., $T_1, T_2, \dots, T_N$. First, each evolutionary process applies the genetic operators described in Section 3.2.1 over a population of individuals. Then, the process presented in Section 3.2.2 makes the population evolve based on the optimization of a fitness function during a specific number of generations.

bigTriGen receives a three-dimensional dataset, D_{3D}, as an input. Each slice of time represents an image of a crop, where each pixel (x,y) is a space point representing the value of a particular vegetation index collected at a given instant $t_i \in \{t_1, t_2, \dots, t_{TP}\}$. Therefore, the D^F set corresponds to the X coordinates of the image and the D^I set to the Y coordinates. Fig. 1a shows the $NDVI$ index represented on the images. That is, the point $(200, 81)$ at t_1 is the $NDVI$ value of the pixel in the row 81 and column 200 at the time instant t_1.

An individual of the evolutionary process corresponds to a tricluster, T_i, being a particular area from the whole input space at a given time window. Thus, T_i is a subset of X and Y coordinates and a continuous subset of time points. Fig. 1b shows an individual represented by the subspace limited by the coordinates $Y = \{87, 88, 89, 90, 91, 92, 93, 94\}$ and $X = \{200, 201, 202, 203, 204\}$, and containing the index values for the time series from t_8 to t_{11}.

The output of the bigTriGen algorithm is a set of triclusters that correspond to a triclustering model $M_{D_{3D}}$ of the input dataset D_{3D}. Each tricluster of this model is a sub-area of an original image of the input dataset, as shown in Fig. 1d. Fig. 1c depicts the behavior patterns for each of the time series that make up a tricluster. Each (x,y) point corresponds to a time series of the specific vegetation index.

Therefore, the aim of the bigTriGen algorithm is to discover a triclustering model, $M_{D_{3D}}$, from a three-dimensional dataset, D_{3D}, where each tricluster determines a sub-area of the original image of the dataset and the time series associated to the tricluster present patterns with similar behavior.

3.2.1. Genetic operators

Several updates have been carried out in bigTriGen with respect to TriGen to deal with satellite imagery related to the precision agriculture. These updates are mainly focused on the genetic operators, which are described below.

Initial population In this phase, the initial individuals of the populations are built. A subset of X and Y coordinates are randomly selected from the input dataset. The (x,y) points resulting from the combination of both subsets is a subspace of the original image of the input dataset. Each new individual's time points are randomly selected from the input dataset, forming a continuous sequence. The number of individuals built in the initial population is determined by the control parameter In.

Selection A tournament algorithm is chosen for this operator. The individuals of the population are firstly separated into three groups, then they are ordered by fitness. A percentage of the population is selected from these three ordered groups. These selected individuals are directly promoted to the next generation and will be the parents suitable for reproduction by applying the crossover operator. The percentage of selected individuals is defined by the control parameter Sel.

Crossover Two individuals are randomly chosen for the reproduction from the individuals selected by the selection operator. From these two parent individuals P_1 and P_2, two new children CH_1 and CH_2 will be obtained as shown in Eqs. (7) and (8). The first new individual CH_1 is composed of the X coordinates of the P_1, the Y coordinates of the P_2, and the time points of the P_1. The second one CH_2 is composed of the X coordinates of the P_2, the Y coordinates of the P_1, and the time points of P_2.

$$P_1 = \left\{P_1^X, P_1^Y, P_1^{TP}\right\} \quad \text{and} \quad P_2 = \left\{P_2^X, P_2^Y, P_2^{TP}\right\} \tag{7}$$

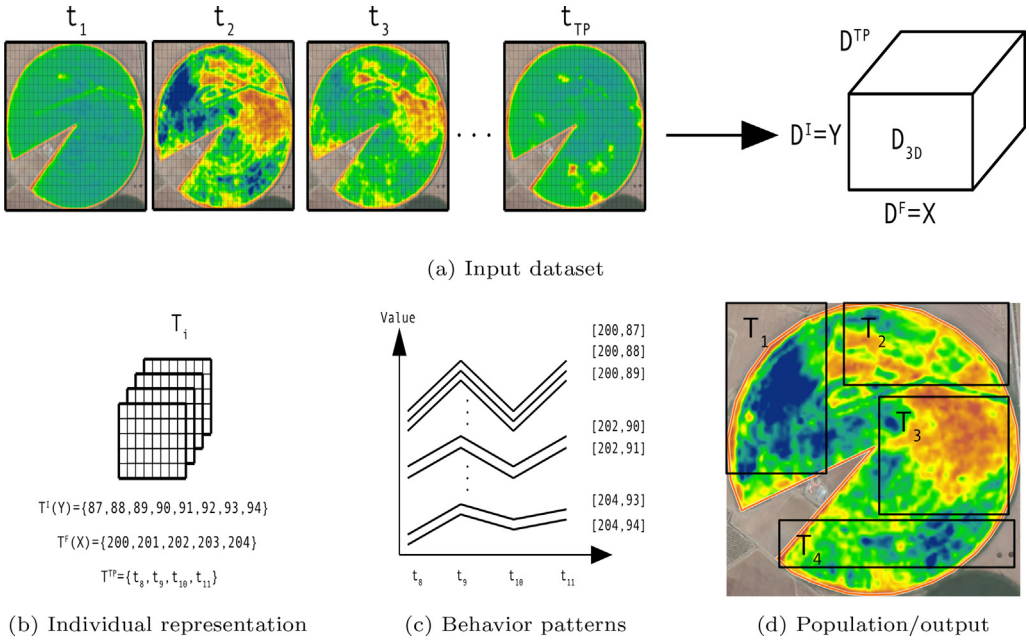
$$CH_1 = \left\{P_1^X, P_2^Y, P_1^{TP}\right\} \quad \text{and} \quad CH_2 = \left\{P_2^X, P_1^Y, P_2^{TP}\right\} \tag{8}$$

The number of children to obtain is $In - (In \times Sel)$, considering Sel as a percentage of selected parents. In order to get them, num_{cross} crossovers are made, where num_{cross} is the quotient plus the remainder of the division by two of the number of children to obtain. As previously mentioned, from one crossover, two children are obtained. Once all crossovers are computed, the specified number of children are selected from all children considering the best fitness function.

Mutation The new individuals obtained by means of the crossover operator are elegible to be altered by mutation. An individual can be altered by removing or adding a random X or Y coordinate or a random time point. The probability of mutation of an individual is set by the control parameter Mut. The operations are controlled by specific parameters referring to maximum and minimum number of coordinates that an individual must have.

3.2.2. Fitness function

As a genetic algorithm, the core of bigTriGen is the fitness function to be optimized. In this work, a fitness function based on the MSL measure [23] has been used. MSL measures the similarity of the behavior patterns contained in a tricluster and it is based on the differences between the angles that every two points of a series form with the X-axis (the slope of a straight line). Thus, this algorithm provides an accurate measure of how similar the behavior



(a) Input dataset

(b) Individual representation

(c) Behavior patterns

(d) Population/output

Fig. 1. TriGen overview.

patterns inside a tricluster are. The MSL measure is widely explained and discussed in [23]. Moreover, the fitness function includes a control mechanism to balance the size of the triclusters and the overlapping among them.

The fitness function of a tricluster T is defined by a weighted average as follows:

$$Fitness(T) = \frac{w_{msl} * MSL(T) + w_s * S(T^X, T^Y) + w_o * O(T, M_{D_{3D}})}{w_{msl} * w_s * w_o} \quad (9)$$

where $MSL(T)$ is the MSL index of the tricluster $T, S(T^X, T^Y)$ is the size of the area demarcated by the X and Y coordinates of the tricluster $T, O(T, M_{D_{3D}})$ is the overlapping degree of the tricluster T with the remaining triclusters of the model $M_{D_{3D}}$, and w_{msl}, w_s and w_o are the weights of each component, respectively [23].

3.3. Big data implementation remarks

The bigTriGen algorithm has been developed in a big data environment to provide it with the ability to analyze big three-dimensional datasets. Therefore, a model with bigger triclusters (more X and Y coordinates and time points) will be discovered from datasets with more significant time points and/or more significant areas (X, Y). bigTriGen has been implemented in Scala 2.12 [32] with Apache Spark 2.3.4 [33]. Its implementation is based on the DataFrame object of Apache Spark. The main feature of this data structure is to be distributed through the nodes of the cluster where the application is deployed [34].

For the bigTriGen algorithm, the input dataset D_{3D} is loaded into a DataFrame, where each row represents a point $(instance, feature, time)$ and its associated value.

The population is also implemented using a DataFrame. An example of the structure can be found in Table 1. In this case, each row represents a time series for the particular coordinates (y, x) of a tricluster individual. Therefore, a row will be composed of an individual numerical identifier (IND_{id}), a time series identifier (TS_{id}), the associated Y and X coordinates, the time point list (TPs) and, the time series values (TS). The justification of these implementation decisions is due to two aspects, both related to the application of the Spark DataFrame API actions and transformations to the population. On the one hand, this structure leads the Spark's actions and transformations to execute the genetic operators in a best-optimized way in a big data environment. On the other hand, this structure boosts the application of the Spark DataFrame API actions and transformations to the population and, therefore, maximizes the distribution of it through the nodes of the Spark cluster where bigTriGen was deployed. In conclusion, with this implementation, the bigTriGen algorithm's scalability, regarding the execution time against the size of the input dataset, is reached. As explained in the above paragraphs, bigTriGen is a novel algorithm with an own design and implementation. A summary of the new features of the bigTriGen is shown in Table 2 where it can be confirmed that bigTriGen differs in the implementation, characteristics and results comparing with TriGen. The original TriGen and the new bigTriGen keep the control parameters of the algorithm, the evolutionary work-flow, and the selection operator in common. In contrast, as discussed above, the bigTriGen allows for the analysis of input datasets and triclusters with sizes impossible to manage on a single machine. Furthermore, it adds the space and time series modeling (presented in Section 3.2) and, therefore, new initial population, crossover and, mutation operators. A detailed description of the original TriGen algorithm can be found in [6,23].

3.4. Validation of the triclusters

In this work, the triclusters of the model $M_{D_{3D}}$ will be validated in three ways. Firstly, the TRIQ quality measure [24] that provides

Table 1
DataFrame example for the population.

IND_{id}	TS_{id}	Y	X	TPs	TS
1	0	2	3	{2,3,4,5}	{0.05,0.58,0.23,0.22}
1	1	2	4	{2,3,4,5}	{0.15,1.82,0.38,0.25}
1	2	3	3	{2,3,4,5}	{0.54,2.84,1.25,0.15}
1	3	3	4	{2,3,4,5}	{0.23,0.38,2.23,1.01}
2	0	20	21	{19,20,21,22,23}	{0.08,0.81,0.09,0.12,2.24}
2	1	20	22	{19,20,21,22,23}	{0.01,1.12,0.01,0.09,1.25}
2	2	21	21	{19,20,21,22,23}	{0.02,1.20,0.02,0.14,3.12}
2	3	21	22	{19,20,21,22,23}	{0.03,1.25,0.25,0.15,5.02}

Table 2
Similarities and differences between TriGen and bigTriGen.

Common features	New features of bigTriGen
Control parameters	Bigger input datasets
	Bigger triclusters
Evolutionary process	(X, Y) space modeling for instances and features
	Time series modeling (consecutive instant points)
Selection operator	Initial population operator
	Crossover operator
	Mutation operator

an index to determine the similarity of the patterns of the triclusters and the correlation level of the time series associated to the tricluster will be used. In particular, TRIQ combines weighted Pearson and Spearman correlation values with a weighted normalization of MSL angle value. It has been shown as a valid measure for representing and summarizing the quality of the triclusters.

Secondly, a visual analysis of the discovered patterns will be carried out. This analysis will determine the coherence of the discovered triclusters in relation to the input dataset. The time series plots will be graphed, and the average of the time series will assess the cohesion of the values of the tricluster. Furthermore, an analysis of the behavior observed will be performed by an expert.

Finally, a global study of the located areas will be also made. The demarcated areas for each tricluster of the model will be also analyzed by an expert. That is necessary to determine any intra-relation between the selected zones for the different triclusters of the model.

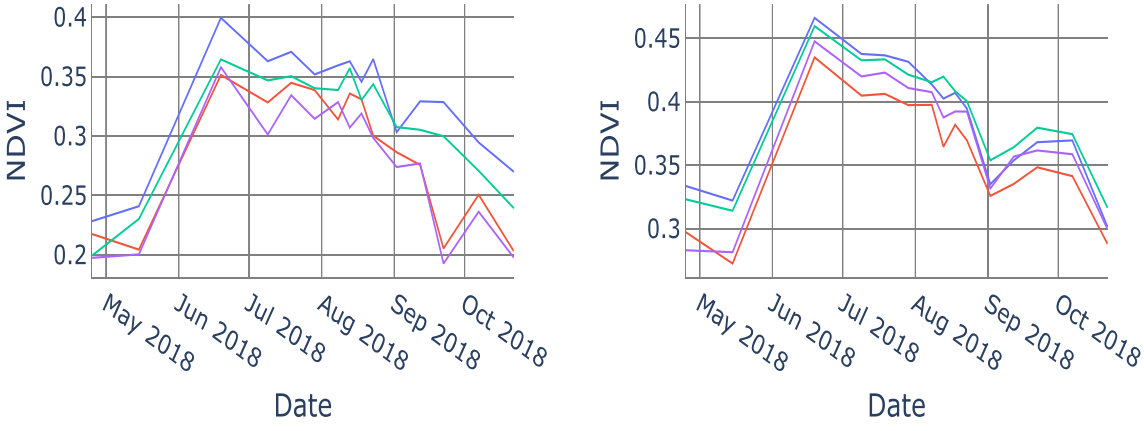
4. Results and discussion

This section reports the analysis of the results obtained by the bigTriGen algorithm when applying to a crop image dataset. In particular, the dataset and the vegetation indices typically used in the crops are described in Section 4.1, the experimentation process is explained in Section 4.2, the discussion of the patterns is presented in Section 4.3 and the scalability analysis to show the ability of the proposed algorithm to deal with big data is in Section 4.4.

4.1. Dataset and vegetation indices

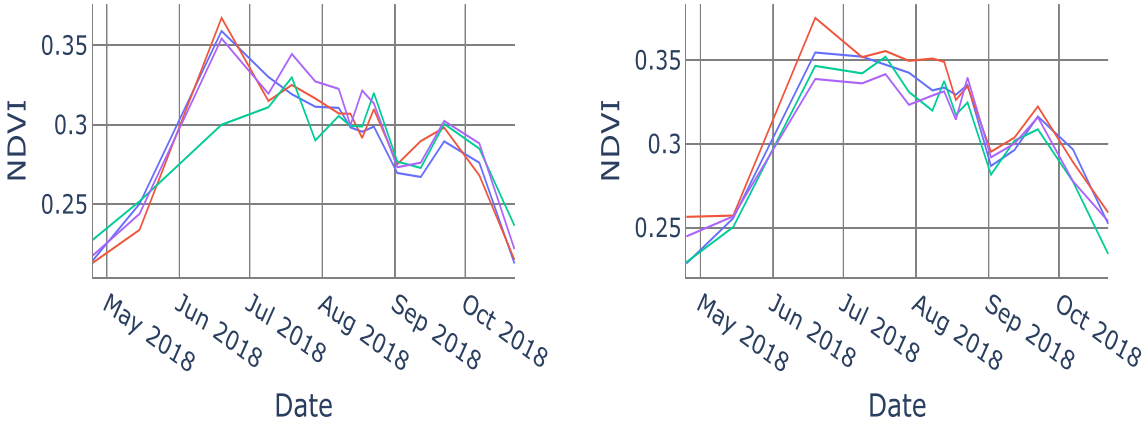
The bigTriGen algorithm is tested in a vineyard crop located in Baixo Alentejo, in Portugal. The study area has 5.15 hectares and its center at the coordinates 37°56'43.62"N 7°52'15.06"W. In particular, the field is monitored during three years (2018, 2019 and 2020) selecting the months that correspond to vineyard season. Data is extracted from Sentinel-2 imagery with high spatial resolution at the defined coordinates using the QGIS software and its Semi-Automatic Classification Plugin. The calculation of the vegetation indices of each image is also made with this software.

Vegetation indices allow the quantitative and qualitative evaluation of different measures of crops, as cover, vigor, growth, type or



(a) NDVI tricluster

(b) NDVI tricluster 2



(c) NDVI tricluster 3

(d) NDVI tricluster 4

Fig. 2. Triclusters using the NDVI index for the vineyard crop.

quality. They are based on the measured canopy reflectance of different wavelength bands [35]. This canopy reflectance can be detected remotely using satellite imagery as the one provided by Sentinel-2. In this particular study, measuring leads to monitor fruit ripening to develop a site-specific harvesting of each zone of the vineyard crop; it is known as Precision Agriculture or more specifically in this case, Precision Viticulture [36].

One of the most used vegetation indices is the NDVI index. This index is very related to the content of the vegetation and varies from 1.0 to −1.0, where 1.0 corresponds to the denser and healthier areas. NDVI includes in its calculation the near-infrared band (*NIR*) and the band for the red (visible) regions (*Red*). NDVI formula is $NDVI = \frac{NIR-Red}{NIR+Red}$. NDVI is a very useful index, for example, to determine areas of a corn crop that behaves differently [20].

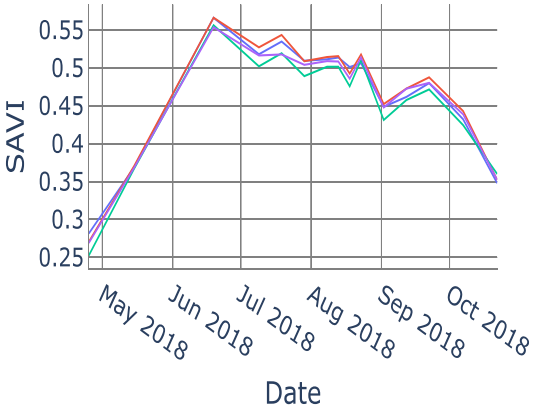
Other indices used in this study that improve NDVI are the Soil-Adjusted Vegetation Index (SAVI) and the Enhanced Vegetation Index (EVI) used to determine grapevine phenology in [37]. The first one introduces L as a correction factor for soil brightness and the second one adds two C_1 and C_2 coefficients to the atmospheric resistance and the *Blue* band, respectively. SAVI is defined as $SAVI = \frac{NIR-Red}{NIR+Red+L} \times (1+L)$ and EVI as $EVI = 2.5 \times \frac{NIR-Red}{NIR+C_1 \times Red - C_2 \times Blue + L}$. In this study, L is 0.5, C_1 is 6 and C_2 is 7.5; they are used values for this kind of crop.

The Moisture Stress Index (MSI) and the Green Normalized Difference Vegetation Index (GNDVI) include two different bands: *Green* and middle-infrared (*MIR*), respectively. GNDVI is sensible to the variation of chlorophyll in the crop. On its side, MSI is used to analyze the water stress and it usually varies from 0.4 to 2 where higher values mean higher water stress and so, less soil moisture. Both indices are, as the above-mentioned ones, very studied in vine crops. For example, [38] concludes their vineyard crop study identifying the MSI as the only vegetation index directly related to the content of the vegetation. GNDVI is represented as $GNDVI = \frac{NIR-Green}{NIR+Green}$ and MSI as $MSI = \frac{MIR}{NIR}$

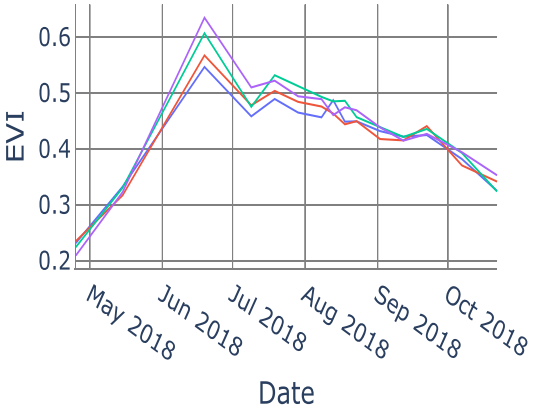
4.2. Experimental setup

The experiments are run on a cluster located at the Data Science and Big Data Laboratory in Pablo de Olavide University. The cluster is made up of four nodes: one master and three slaves. It has four Processors Intel(R) Core (TM) i7-5820 K CPU with 48 cores, 120 GB of RAM memory. The cluster uses Ubuntu 16.04 LTS, Apache Spark 2.3.4 and HDFS file system on Hadoop 2.7.7. The bigTriGen algorithm is implemented in Scala programming language.

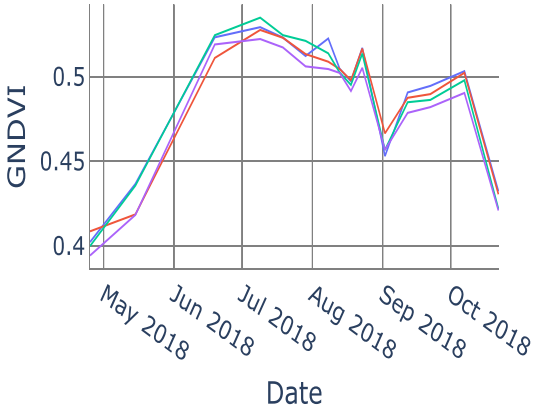
Considering the work published in [20] that discovered three dimensional patterns in a maize plantation area in Baixo Alentejo and after several experimental tests with the bigTriGen algorithm, the selected control parameters for the experimentation are: N = 4, G = 10, In = 200, Sel = 0.8 and Mut = 0.1. The fitness function used



(a) SAVI tricluster



(b) EVI tricluster



(c) GNDVI tricluster

Fig. 3. Triclusters using SAVI, EVI and GNDVI for the vineyard crop.

is described in Section 3.2.2 and the validation of the triclusters is made considering the remarks in Section 3.4.

4.3. Pattern discovery

The process of discovering three-dimensional patterns using the proposed algorithm is performed from two points of view: spatial and temporal.

4.3.1. Spatial patterns

The goal of this analysis is to find behavior patterns that identify spatial zones on the vineyard crop with different characteristics. This analysis is carried out for the 2018 growing season.

Fig. 2 represents the triclusters found by the bigTriGen algorithm using the NDVI index. It shows a great uniformity between all the areas of each sub-figure, as the discovered patterns show very similar behavior curves. In order to confirm this uniformity, more vegetation indices that consider corrections of the NDVI and more environmental factors are used.

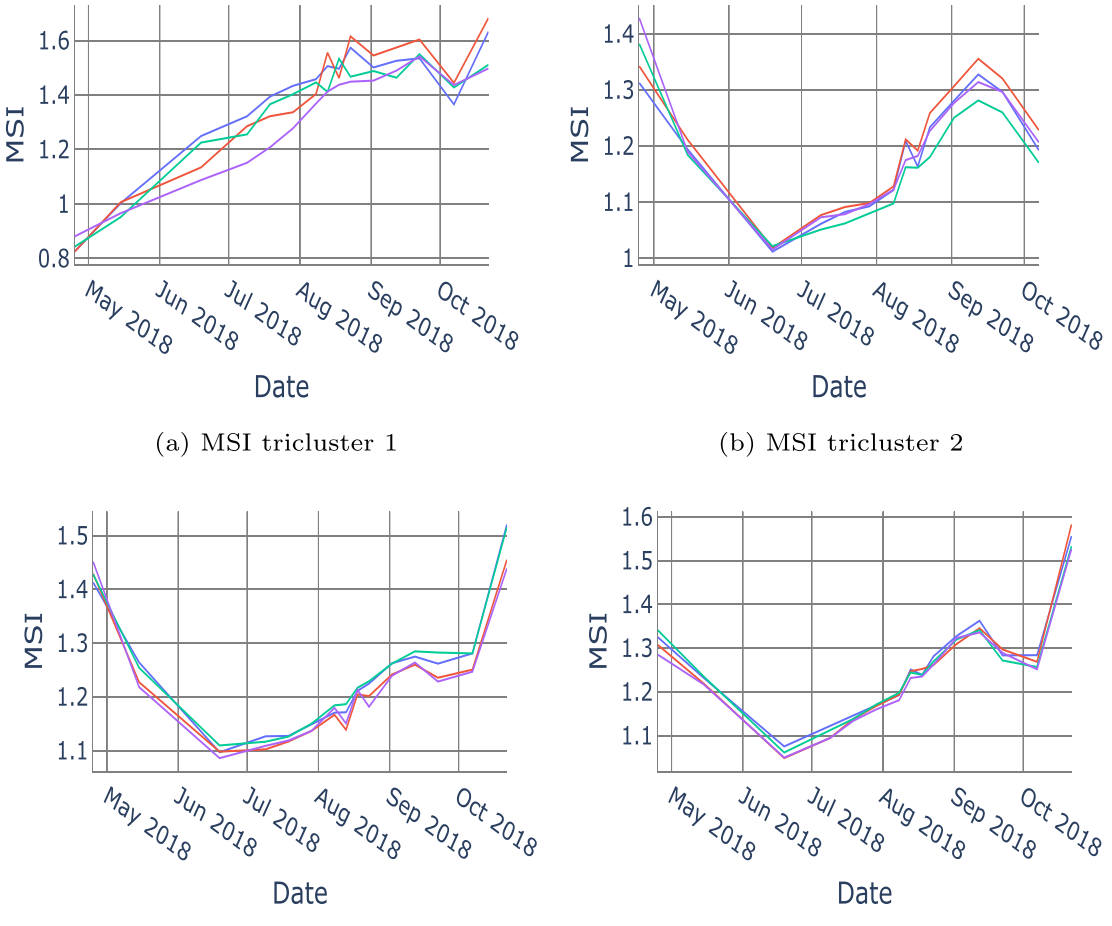
Fig. 3 depicts the behavior patterns of the field with SAVI, EVI and GNDVI indices during 2018. The analyses carried out using these indices confirm the assessment made with the NDVI index, i.e., triclusters curves represent an uniform behavior for vegetative growth and development of the crop throughout its extension.

To further study the water stress to which the crop is subjected, an additional analysis is carried out with the MSI index, which introduces *MIR* band in its calculation. Areas of the crop with different trends in the value of water stress are identified, although this stress does not imply effects on the crop that are perceptible when the rest of indices are used.

Fig. 4 illustrates the different behavior patterns of the four triclusters obtained by the proposed bigTriGen algorithm when using the MSI. Fig. 4a represents, unlike the other three, an area with higher soil moisture during the initial period of the growing season. The trend is similar in the other behavior patterns, tending towards an increase in water stress as the growing season progresses. However, in the final phase, close to harvest time, is where the greatest differences among the different triclusters identified can be seen. While water stress is maintained in the area represented in Fig. 4a, a clear increase in stress is observed in Fig. 4c and d, in contrast to an increase in soil moisture in Fig. 4b, considering that lower MSI values correspond to lower water stress and so, higher soil moisture or water content. The quality of the found triclusters has been measured with the TRIQ measure described in Section 3.4. For values that move in the [0–1] interval, the first tricluster has a TRIQ value of 0.8799, the second of 0.9365, the third of 0.9153 and the forth of 0.8321, thus ensuring accurate patterns for all cases.

This information may be relevant for the analysis of productivity in each field zone. It is necessary to identify the causes of the different behavior in order to determine whether they are due to productive factors (irrigation inequality, pests, etc.) or to specific factors of the terrain (inclines, type and quality of the terrain, etc.).

Fig. 5 identifies the geographic areas in the field map represented by the found triclusters when using the MSI index.



(a) MSI tricluster 1

(b) MSI tricluster 2

(c) MSI tricluster 3

(d) MSI tricluster 4

Fig. 4. Triclusters using the MSI for the vineyard crop.

Fig. 5. Geographic location of the triclusters using the MSI index in the vineyard crop.

4.3.2. Temporal patterns

In this Section, the search of different behavior patterns between different growing seasons is considered. Years 2018, 2019 and 2020 have been analyzed, in particular, the analyses were limited to the months between May and November, which correspond to the vine growing season in its different phases.

The results obtained from the determination of triclusters for the NDVI data in the different growing seasons confirm the uniformity of the patterns in terms of crop behavior, i.e., patterns identified each year are very similar to the others found in the same year. However, there is a clear difference between one year and another. The patterns of 2018 are in Fig. 2 and a representation of the ones of 2019 and 2020 in Fig. 6.

The patterns of the last months of the vineyard period for the years 2018 and 2020 are very similar and correspond to the theory of what the NDVI trend should be over the course of a growing season. Nevertheless, the triclusters for 2019 show a different behavior during the month of August. In that period, the crop suffered a drop in NDVI index indicating a loss of quality of the plantation, which has managed to recover in the following months. This incidence coincides with the period of severe forest fires in the area where the field is located. It is very likely that this is the cause of the temporary deterioration of the crop.

The bigTriGen algorithm demonstrates with this analysis that it is suitable for discovering anomalous behavior in a temporal sequence of historical events. Its use with current data can be a good tool to detect indications of anomalies at an early stage, even not perceptible to the naked eye in the crop, allowing corrective measures to be taken as soon as possible to mitigate the effects of these occurrences.

4.4. Scalability analysis

Once the found patterns have been evaluated, the next step is to study the scalability of the proposed bigTriGen algorithm. The evolution of the execution times is analyzed in two parts: in the first one, considering the effect of the number of nodes used and in the second one, considering the influence of the size of the dataset used. These tests are executed with a base dataset with the same characteristics as the one defined in Section 4.1 and the optimal parameters described in Section 4.2.

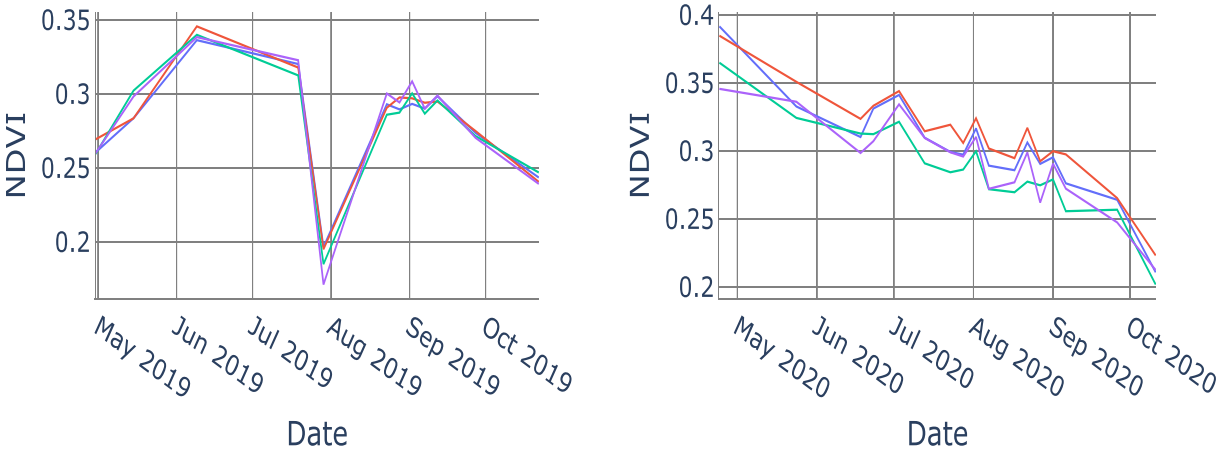
First, the scalability in terms of resources is analyzed by changing the number of nodes used when executing the bigTriGen algorithm. As explained in Section 4.2, the cluster used is made up of four nodes with twelve cores in each node. This analysis is made with 12 cores, 24 cores, 36 cores and 48 cores.

To analyze the effects of the dataset size, a base dataset of a size of 65 MiB has been used. This characteristic of the scalability analysis is studied by multiplying the length of the base dataset by 1, 2, 4, 6, 16 and 32. It corresponds to six experiments with datasets of 65 MiB, 130 MiB, 260 MiB, 520 MiB, 1040 MiB and 2080 MiB, respectively.

Results of twenty-four scalability experiments are shown in Table 3 and in Fig. 7, where the execution times are presented in

Table 3
Execution times (in minutes) according to number of cores and size of datasets.

Multiplier	12 cores	24 cores	36 cores	48 cores
x 1	19.7759	19.6475	20.3361	20.8268
x 2	25.1146	26.7812	25.9973	26.5552
x 4	42.2082	39.5204	38.5376	40.4623
x 8	76.0349	68.0430	69.9858	64.0017
x 16	141.9072	120.7762	120.9260	125.3108
x 32	278.6696	270.0996	232.3536	239.3660



(a) NDVI tricluster in 2019 (b) NDVI tricluster in 2020

Fig. 6. Triclusters using the NDVI in the years 2019 and 2020.

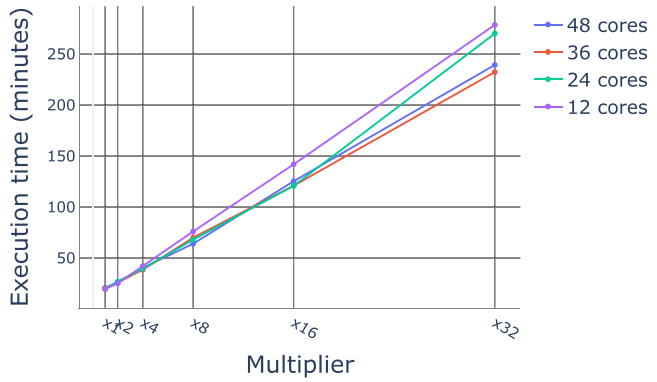


Fig. 7. Scalability analysis.

minutes. The computing time is not very influenced by the number of cores used when the size of the dataset is small, i.e., x1, x2, x4 or x8, but when the size increases, the execution time is smaller for 36 and 48 cores.

The behavior of the bigTriGen leads to express its scalability factor as $Factor_i = \frac{size_i}{size_{\frac{i}{2}}}$, where *size* represents the dataset size and i varies from 2 to 32. This factor is usually smaller than 2 which is better than linear scalability. It is important to considering that the bigTriGen algorithm is influenced by chance, for example by means of the mutation operation, among others. However, in order to get a comparable scalability analysis, these operators have been controlled.

5. Conclusions

In this paper the new bigTriGen triclustering algorithm has been introduced to mine three-dimensional patterns from big datasets. In particular, this algorithm has used specific genetic operators to find triclusters in addition to control the overlapping with the previously found tricluster solutions. The bigTriGen has been applied to a vineyard crop in southern Portugal to find a precision viticulture solution. The accuracy of the algorithm has been shown with respect to two different features: the quality measure of the found patterns and the scalability of the algorithm. On the one hand, different vegetation indices have been calculated using Sentinel-2 images downloaded from QGIS software. The found patterns using these vegetation indices have shown that the index that best fits this field is the MSI. In this way, the algorithm has been able to find four different areas of the vineyard crop that behave differently in terms of their soil moisture. In addition, the algorithm has found different behaviors of the crop during 2018, 2019 and 2020. On the other hand, the scalability of the algorithm has been studied considering the number of nodes used and the size of the dataset. In both cases, the scalability factor of the bigTriGen has been proven to be even better than linear scalability.

The future works will be focused on developing more characteristics of the algorithm such as detecting anomalies in streams, creating methods to select the optimal values of the parameters or using other type of data.

CRediT authorship contribution statement

Laura Melgar-García: Validation. **David Gutiérrez-Avilés:** Conceptualization, Methodology. **Maria Teresa Godinho:** Writing – review & editing. **Rita Espada:** Data curation, Validation. **Isabel Sofia Brito:** Visualization, Investigation. **Francisco Martínez-Álvarez:** Supervision, Investigation. **Alicia Troncoso:** Supervision. **Cristina Rubio-Escudero:** Conceptualization, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank the Spanish Ministry of Science and Innovation for the support under the project PID2020-117954RB and the European Regional Development Fund and Junta de Andalucía for projects PY20-00870 and UPO-138516. The authors also thank the Portuguese Agency "Fundação para a Ciência e a Tecnologia" (FCT), in the framework of the project UIDB/00066/2020. This work could not have been done without the support and help of the Farmer's Association of Baixo Alentejo and Francisco Palma during the whole project. Finally, the authors thank António Vieira Lima and Moragri S. A. for giving access to data.

References

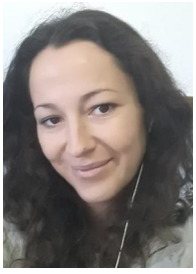
[1] N. Khan, I. Yaqoob, I.A.T. Hashem, Z. Inayat, W.K. Mahmoud Ali, M. Alam, M. Shiraz, A. Gani, Big Data: Survey, Technologies, Opportunities, and Challenges, Scientific World J. 2014 (2014) 712826.
[2] A. Galicia, J.F. Torres, F. Martínez-Álvarez, A. Troncoso, A novel spark-based multi-step forecasting algorithm for big data time series, Inf. Sci. 467 (2018) 800–818.
[3] C.C. Aggarwal, Data Mining: The Textbook, Springer Publishing Company, Incorporated, 2015.
[4] F.J. Pierce, P. Nowak, Aspects of precision agriculture, Adv. Agron. 67 (1999) 1–85.
[5] J. Tan, P. Yang, Z. Liu, W. Wu, L. Zhang, Z. Li, L. You, H. Tang, Z. Li, 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) (2014) 397–410.
[6] D. Gutiérrez-Avilés, C. Rubio-Escudero, F. Martínez-Álvarez, J. Riquelme, TriGen: A genetic algorithm to mine triclusters in temporal gene expression data, Neurocomputing 132 (2014) 42–53.
[7] L. Santesteban, S. Guillaume, J. Royo, B. Tisseyre, Are precision agriculture tools and methods relevant at the whole-vineyard scale?, Precision Agric 14 (2012) 2–17.
[8] R. Plant, Site-specific management: The application of information technology to crop production, Comput. Electron. Agric. 30 (2001) 9–29.
[9] J. Costa, M. Vaz, J. Escalona, R. Egipto, C. Lopes, H. Medrano, M. Chaves, Modern viticulture in southern Europe: Vulnerabilities and strategies for adaptation to water scarcity, Agric. Water Manag. 164 (2016) 5–18.
[10] A. Khaliq, L. Comba, A. Biglia, D. Ricauda Aimonino, M. Chiaberge, P. Gay, Comparison of Satellite and UAV-Based Multispectral Imagery for Vineyard Variability Assessment, Remote Sens. 11(4)..
[11] A. Matese, P. Toscano, S.F. Di Gennaro, L. Genesio, F.P. Vaccari, J. Primicerio, C. Belli, A. Zaldei, R. Bianconi, B. Gioli, Intercomparison of UAV, Aircraft and Satellite Remote Sensing Platforms for Precision Viticulture, Remote Sens. 7 (3) (2015) 2971–2990.
[12] S.F. Di Gennaro, R. Dainelli, A. Palliotti, P. Toscano, A. Matese, Sentinel-2 Validation for Spatial Variability Assessment in Overhead Trellis System Viticulture Versus UAV and Agronomic Data, Remote Sens. 11(21)..
[13] L. Pastonchi, S. Di Gennaro, P. Toscano, A. Matese, Comparison between satellite and ground data with uav-based information to analyse vineyard spatio-temporal variability, XIIIth International Terroir Congress, OENO One 54 (2020) 919–934.
[14] C. von Hebel, S. Reynaert, K. Pauly, P. Janssens, I. Piccard, J. Vanderborght, J. Kruk, H. Vereecken, S. Garre, Toward high-resolution agronomic soil information and management zones delineated by ground-based electromagnetic induction and aerial drone data, Vadose Zone J..
[15] P. Janrao, H. Palivela, Management zone delineation in Precision agriculture using data mining: A review, in: 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015, pp. 1–7..
[16] B.I. Evstatiev, K.G. Gabrovska-Evstatieva, A review on the methods for big data analysis in agriculture, IOP Conference Series: Materials Science and Engineering 1032 (2021) 012053.
[17] A. Gavioli, E.G. de Souza, C.L. Bazzi, L.P.C. Guedes, K. Schenatto, Optimization of management zone delineation by using spatial principal components, Comput. Electron. Agric. 127 (2016) 302–310.

[18] P. Cinat, S.F. Di Gennaro, A. Berton, A. Matese, Comparison of Unsupervised Algorithms for Vineyard Canopy Segmentation from UAV Multispectral Images, Remote Sens. 11(9)..
[19] N. Ohana-Levi, K. Knipper, W.P. Kustas, M.C. Anderson, Y. Netzer, F. Gao, M. d. M. Alsina, L.A. Sanchez, A. Karnieli, Using Satellite Thermal-Based Evapotranspiration Time Series for Defining Management Zones and Spatial Association to Local Attributes in a Vineyard, Remote Sensing 12 (15)..
[20] L. Melgar-García, M.T. Godinho, R. Espada, D. Gutiérrez-Avilés, I.S. Brito, F. Martínez-Álvarez, A. Troncoso, C. Rubio-Escudero, Discovering spatio-temporal patterns in precision agriculture based on triclustering, in: 15th International Conference on Soft Computing Models in Industrial and Environmental Applications, Springer, Cham, 2021, pp. 226–236.
[21] R.U.I. Henriques, S.C. Madeira, Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey, ACM Comput. Surv. 51 (5) (2018) 43.
[22] D. Gutiérrez-Avilés, C. Rubio-Escudero, Mining 3D patterns from gene expression temporal data: A new tricluster evaluation measure, Scientific World J. 2014 (2014) 1–16.
[23] D. Gutiérrez-Avilés, C. Rubio-Escudero, MSL: A measure to evaluate three-dimensional patterns in gene expression data, Evol. Bioinformatics 11 (2015) 121–135.
[24] D. Gutiérrez-Avilés, R. Giráldez, F.J. Gil-Cumbreras, C. Rubio-Escudero, TRIQ: a new method to evaluate triclusters, BioData Mining 11 (2018) 15.
[25] D. Gutiérrez-Avilés, C. Rubio-Escudero, LSL: A new measure to evaluate triclusters, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 2014, pp. 30–37..
[26] F. Martínez-Álvarez, G. Asencio-Cortés, J.F. Torres, D. Gutiérrez-Avilés, L. Melgar-García, R. Pérez-Chacón, C. Rubio-Escudero, J.C. Riquelme, A. Troncoso, Coronavirus Optimization Algorithm: A Bioinspired Metaheuristic Based on the COVID-19 Propagation Model, Big Data 8 (4) (2020) 308–322.
[27] L. Melgar-García, D. Gutiérrez-Avilés, C. Rubio-Escudero, A. Troncoso, High-content screening images streaming analysis using the strigen methodology, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, Association for Computing Machinery, 2020, pp. 537–539.
[28] F. Martínez-Álvarez, D. Gutiérrez-Avilés, A. Morales-Esteban, J. Reyes, J.L. Amaro-Mellado, C. Rubio-Escudero, A novel method for seismogenic zoning based on triclustering: Application to the Iberian Peninsula, Entropy 17 (7) (2015) 5000–5021.
[29] L. Melgar-García, D. Gutiérrez-Avilés, C. Rubio-Escudero, A. Troncoso, Discovering three-dimensional patterns in real-time from data streams: An online triclustering approach, Inf. Sci. 558 (2021) 174–193.
[30] R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, Parallel batch k-means for big data clustering, Comput. Ind. Eng. 152 (2021) 107023.
[31] C.-E. Ben Ncir, A. Hamza, W. Bouaguel, Parallel and scalable dunn index for the validation of big data clusters, Parallel Comput. 102 (2021) 102751.
[32] M. Odersky, L. Spoon, B. Venners, Programming in Scala: Updated for Scala 2.12, third ed., Artima Incorporation, Sunnyvale, CA, USA, 2016.
[33] B. Chambers, M. Zaharia, Spark: The Definitive Guide Big Data Processing Made Simple, first ed., O'Reilly Media Inc, 2018.
[34] M. Zaharia, R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M.J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache Spark: A Unified Engine for Big Data Processing, Commun. ACM 59 (11) (2016) 56–65.
[35] X. Jinru, B. Su, Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications, J. Sens. 2017 (2017) 1–17.
[36] A. Martínez, V.D. Gomez-Miguel, Vegetation index cartography as a methodology complement to the terroir zoning for its use in precision viticulture, OENO One 51 (3) (2017) 289.
[37] H. Fraga, M. Amraoui, A. Malheiro, J. Moutinho Pereira, J. Eiras-Dias, J. Silvestre, J. Santos, Examining the relationship between the enhanced vegetation index and grapevine phenology, Eur. J. Remote Sens. 47 (2014) 753–771.
[38] E. Laroche-Pinel, M. Albughdadi, S. Duthoit, V. Chéret, J. Rousseau, H. Clenet, Understanding vine hyperspectral signature through different irrigation plans: A first step to monitor vineyard water status, Remote Sens. 13 (3) (2021) 31.

Laura Melgar-García is a PhD student in Computer Science at the Pablo de Olavide University with a pre-doctoral researcher grant (FPU) from the Ministry of Science and Innovation of Spain. Before starting her doctoral studies, she earned a Biomedical Engineering Degree in 2017 and a Master in Software Engineering in 2018. Her major fields of research are the modeling and analysis of massive data, focusing on batch and streaming/online processing. The object of her research seeks to obtain both descriptive and predictive models with applications in real problems as precision agriculture, energy consumption or biomedical solutions.

David Gutiérrez-Avilés received the PhD degree in computer engineering from the University of Seville. He has been with the Department of Computer Science at the Pablo de Olavide University since 2016, where he is currently an assistant teacher. His primary areas of interest are bioinformatics, machine learning, data mining, and big data analytics.

Maria Teresa Godinho received the PhD degree in Operations Research from the University of Lisbon. She has been with the Department of Mathematical and Physical Sciences at the Instituto Politecnico de Beja, Portugal, since 2001, where she holds a position as an Adjunct Professor. Her primary areas of interest are integer programming and optimization algorithms.



Rita Isabel Espada has a degree and a master's degree in Agronomy from Escola Superior Agraria de Beja. She has been with AABA- Alentejo Farmers Association in Beja, Portugal, since 2018, where I am a technician in integrated production. Where I am responsible for agricultural advice and support for farmers. She has a special interest in precision farming and new technologies.

Isabel Sofia Brito is a Coordinator Professor at Polytechnic Institute of Beja, Portugal, and a member of the Centre of Technology and Systems (CTS-UNINOVA). Her main research interests are Requirements Engineering and Sustainability Requirements, Model and Data-Driven Development, Multi-Criteria Decision Making and, Big Data where she has published several papers on these topics in journals, international and national conferences, and workshops.

Francisco Martínez-Álvarez received the MSc degree in telecommunications engineering from the University of Seville, and the PhD degree in computer engineering from the Pablo de Olavide University. He has been with the Department of Computer Science at the Pablo de Olavide University since 2007, where he is currently a full professor. His primary areas of interest are time series analysis, data mining, and big data analytics.

Cristina Rubio-Escudero received the Ph.D. degree in Computer Science from the University of Granada, Spain. She has been with the Department of Computer Science at the University of Seville since 2007, where she is currently an associate professor. Her primary areas of interest are bioinformatics, machine learning and big data.

Alicia Troncoso received the Ph.D. degree in Computer Science from the University of Seville, Spain, in 2005. She has been with the Department of Computer Science at the Pablo de Olavide University since 2005, where she is currently a full professor. Her primary areas of interest are time series forecasting, machine learning and big data.