Intelligent Environments 2020
C.A. Iglesias et al. (Eds.)

doi:10.3233/AISE200035

Towards the Characterization of Agricultural Regions Based on Weather Conditions - Sustainable Agriculture

Jose M. CADENAS [a,1], M. Carmen GARRIDO [a] and Raquel MARTÍNEZ [b]
[a] *Dept. of Information and Communication Engineering. University of Murcia. Spain*
[b] *Dept. of Computer Engineering, Catholic University of Murcia. Spain*

Abstract. Nowadays, precision agriculture has acquired an important role in improving the existing problems in this domain, allowing to increase agricultural production and, therefore, the economic profits derived from it. The general objective of precision agriculture is to help and/or improve decision making regarding the quality, productivity, profitability, resource efficiency and sustainability of agricultural production. One way to carry out this objective is to analyze how suitable a type of crop is in the geographical area in which it is located. In this work, a Intelligent Data Analysis process is used and the study that should lead us to obtain knowledge of geographic regions with similar characteristics from climate viewpoint is started. This should allow us to advise farmers in areas with weather conditions that will cause a lower yields, which types of crops will lead to higher yields and could be well suited to their areas because they are getting good yield in zones with similar weather conditions.

Keywords. Precision Agriculture, Intelligent Data Analysis, Sustainable Agriculture, Clustering Analysis

1. Introduction

In the agricultural sector, numerous decisions are made every day with the aim of obtaining the best possible yield from crops, both in terms of productivity and the resources necessary to obtain a good production. Nowadays, it is important to support this decision making in the large amounts of data that are being stored from the agricultural environment. This general objective can be achieved by carrying out the grouping of regions with similar weather conditions and then making a yield analysis of each region to advise the farmer of the regions with the worst yield, the kind of cultivation of the regions with a more optimal yield. This can be done by applying an Intelligent Data Analysis process (IDA process) to the large amounts of available data from crops, soils and weather stations that allow a non-experimental data analysis to optimize production and make agriculture more resilient to climate change.

In this work, groupings of regions from different areas of the Region of Murcia (south-east of Spain) are found by means of a clustering technique from the data collected

[1] Corresponding Author, E-mail: jcadenas@um.es

from various weather stations. With the preliminary results obtained in this work, in subsequent works an improvement of the IDA process will be carried out, transferring it to the Softcomputing framework that will allow us to deal with imperfect data and, therefore, to express in a more correct way the true nature of the data.

The structure of the work is as follows: In Section 2 works that try to solve problems in agriculture using clustering techniques are review. In Section 3 the clustering technique used to find the region groupings is introduced. In Section 4 the description of both the data used and the results obtained when applying the technique are shown. In Section 5 the different future works to be carried out based on the preliminary results obtained in this work are indicated. Finally Section 6 shows the conclusions of the results obtained.

2. Clustering in precision agriculture problems

Precision agriculture addresses a wide range of agricultural problems with the aim of making crops more sustainable, enabling farmers to maximise their profits and reduce their losses. One of the tools that precision agriculture uses to predict and help to make decisions is machine learning techniques [1]. This work focuses on clustering techniques mainly, with the aim of grouping and differentiating agricultural areas to address problems such as irrigation or inclement weather, among others.

The most common problem faced by the farmers is they do not opt crop based on the necessity of soil and weather conditions, as a result they face serious setback in productivity. This problem can be addressed through precision agriculture. This strategy takes into account several parameters, viz: soil characteristics and types, weather conditions and crop yield. Data collection based on these parameters suggesting the farmer suitable crop to be cultivated. Precision agriculture helps in reduction of non suitable crop which indeed increases productivity, apart from the following advantages like efficacy in input as well as output and better decision making for farming. In this framework, different contributions have been made in recent years. For example, the spatial-temporal change in agricultural distribution in Thailand is analyzed in [2] during the period 2007-2015 using cluster, outlier and hot spots analysis. The conclusions, and main objectives of the analysis, are to support and contribute to the strengthening of energy and food security through adaptation or survival to climate change for the period 2015-2021.

The authors of [3] address the problem of Indian farmers who do not choose to cultivate according to the need for the soil, and therefore face a serious decline in productivity. They propose a system of recommendation through a majority voting assembly model using techniques such as random tree, k-neighbor and naive bayes to recommend a suitable crop.

In [4], the aim is to predict the rain periods for agricultural applications. This method joins years with similar pluviometric characteristics, with the purpose of finding early a pattern that can define the behavior of the distribution of rainfall between the current year. The method was used operationally during the ENSO phenomenon of 1997-1998, and insubordinated to climatological studies. These have allowed different groups of growers to make climate forecasts for the development of their activities.

In [5], a general study and a comparative study on different types of clusters is carried out to group different cultivation areas. Specifically, the evaluation was carried out with data obtained between 2010 and 2015 from three commercial agricultural elders cultivated with soybean and maize in Brazil. In general, the behaviour of all clustering algorithms was satisfactory for the purpose indicated.

3. IDA process based on a clustering technique

Cluster analysis is one of the most outstanding descriptive tasks in the IDA process. The idea is to partition a dataset into groups with similar characteristics. It is an unsupervised task and the groups obtained could be considered as classes.

In this work is proposed the use of the CMEANS clustering algorithm [6] with a heterogeneous distance measure that will allow us working with the nominal and numeric attributes described. In the next Subsection the CMEANS algorithm is described.

3.1. CMEANS Algorithm

A clustering algorithm groups a dataset E into c partitions trying to keep the data within the same cluster as close as possible and the clusters as far as possible. The objective function plays an important role in a clustering algorithm. Given a dataset $E = \{x^1, x^2, \ldots\}$, where each instance is described by n attributes $x^i = \{x^i_1, x^i_2, \ldots, x^i_n\}$, in the CMEANS Algorithm the objective function to be minimized is defined using the Euclidean distance $J_{CMEANS} = \sum_{j=1}^{|E|} \sum_{i=1}^{c} d(x^j, v^i)$ where c is the number of clusters, $|E|$ is the number of available instances, $v^i = (v^i_1, v^i_2, \ldots, v^i_n)$ is the cluster center i, and $d(x^j, v^i)$ is a function measuring the distance between the instance x^j and the cluster center v^i.

The main steps are shown in the Algorithm 1.

Algorithm 1. CMEANS Algorithm

Input Dataset E, Value c; $1 \leq c \leq |E|$

Initialize randomly the cluster centers vector v

while $v^i(t) - v^i(t-1) > \varepsilon$ **do**

Calculate the index sets $I^i; i = 1, ..., c$ composed with the set of instance indexes that are closer to the cluster center v^i than to any of the other cluster centers.

Recalculate the cluster centers according to $v^i = \frac{\sum_{j \in I^i} x^j}{|I^i|}$

end while

As it was previously commented, in this work a heterogeneous distance function is used because the instances can be described by attributes with nominal and numeric values and the cluster centers too, $F(x^j, v^i) = \frac{\sum_{k=1}^{n} f(x^j_k, v^i_k):}{n}$ where:

- If attribute k is nominal: $f(x^j_k, v^i_k) = 1 - \frac{Card(x^j_k \bigcap v^i_k)}{Card(x^j_k \bigcup v^i_k)}$ where e^j_k, v^i_k are crisp/fuzzy nominal values and $Card(x^j_k \bigcap v^i_k)$ and $Card(x^j_k \bigcup v^i_k)$ are defined as the cardinality crisp subsets obtained by the union and intersection of x^j_k and v^i_k, respectively.
- If attribute k is numeric: $f(x^j_k, v^i_k) = |x^j_k - v^i_k|$, where x^j_k and v^j_k are numeric values.

The update of the cluster centers from the crisp partition I_i is carried out as follows, where $|I_i|$ is the number of instances belonging to cluster i:

If attribute k is nominal: $v_k^i = \frac{1}{|I_i|} \bigcup_{x^j \in I_i} x_k^j$ or, If it is numeric: $v_k^i = \frac{1}{|I_i|} \sum_{x^j \in I_i} x_k^j$

4. Experiments, analysis and discussion

4.1. Information collection and data preparation

This study analyses the Autonomous Community of the Region of Murcia (Spain). The used information is obtained from the Agricultural Information Service (SIAM, http://siam.imida.es) which depends on the Murcian Institute of Agricultural and Food Research and Development. The Region covers the following areas and municipalities:

- Altiplano – Yecla, Jumilla, Abanilla and Fortuna.
- Noroeste – Moratalla, Caravaca de la Cruz, Cehegín and Bullas.
- Río Mula – Mula, Pliego, Albudeite and Campos del Río.
- Vega del Segura – Murcia, Beniel, Santomera, Alcantarilla, Molina de Segura, Torres de Cotillas, Alguazas, Ceutí, Lorquí, Archena, Ulea, Villanueva del Segura, Ojós, Ricote, Blanca, Abarán, Cieza and Calasparra.
- Valle del Guadalentín – Lorca, Puerto Lumbreras, Águilas, Mazarrón, Totana, Aledo, Alhama de Murcia and Librilla.
- Campo de Cartagena – Fuente Álamo, Cartagena, Unión (La), Torre Pacheco, San Javier, San Pedro del Pinatar and Alcázares (Los).

Many of these areas focus on agriculture (fruit and vegetables), which represents a strategic sector in the economy of this Region [7]. A crucial aspect in this economy is the agro-food production for export, since it represents a percentage of more than 50% of the regional market. Of the total area of the region, 50% is used for cultivation (67% as dry land and 33% as irrigated land). Figure 1 shows the regional agricultural production differentiating by percentage each kind of crop.

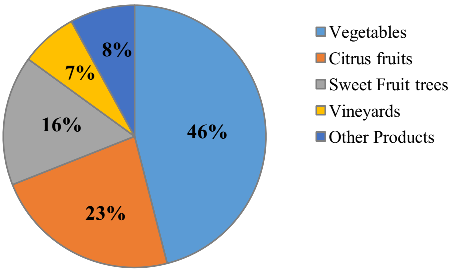
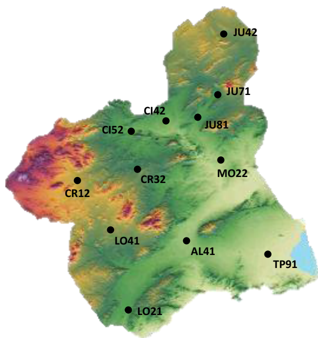


Figure 1. Regional agricultural production of different types of crops in the Region of Murcia, Spain

The data collected correspond to 12 weather stations (see Table 1). Each station is equipped with the following sensors and ephemeris: weather vane, radiometer, rain gauge, data-logger and thermo-hygrometer. The information collected corresponds to 4 years during the periods from 01/12/2016 to 31/03/2019 (approximately the meteorological winter). The initial data obtained from SIAM sensors correspond to values obtained every 5 minutes and these are grouped 12 by 12 to show only values for each hour. For this reason, some of the measurements show the minimum, mean and maximum values for each hour. The type of information obtained is shown in the Table 2. It is important to clarify two measures that have no units, specifically the measure Sunlight is a boolean that indicates whether there is sun or not, while Cooling units is a measure calculated to indicate the hours of cold that support/need crops.

Table 1. Description of the weather stations in study of the Region of Murcia, Spain.



Station CR12 (Noroeste) – Altitude 869m	Coordinate: 38^o 2’ 38.24” – 1^o 58’ 48,67”
Station CR32 (Noroeste)– Altitude 433m	Coordinate: 38^o 6’ 39” – 1^o 19’ 27.58”
Station JU71 (Altiplano) – Altitude 401m	Coordinate: 38^o 23’ 40,01” – 1^o 14’ 21,58”
Station JU81 (Altiplano) – Altitude 341m	Coordinate: 38^o 19’ 11,3” – 1^o 19’ 27,58”
Station JU42 (Altiplano) – Altitude 658m	Coordinate: 38^o 39’ 31,9” – 1^o 11’ 8,73”
Station CI52 (V. Segura) – Altitude 275m	Coordinate: 38^o 15’ 12.59” – 1^o 41’ 41,89”
Station CI42 (V. Segura) – Altitude 244m	Coordinate: 38^o 17’ 2” – 1^o 29’ 46,84”
Station MO22 (V. Segura) – Altitude 146m	Coordinate: 38^o 7’ 39,04” – 1^o 13’ 14,36”
Station TP91 (C. Cartagena) – Altitude 56m	Coordinate: 37^o 44’ 51,81” – 0^o 59’ 12,02”
Station AL41 (V. Guadalentín) – Altitude 169m	Coordinate: 37^o 47’ 32,05” – 1^o 25’ 39”
Station LO21 (V. Guadalentín) – Altitude 356m	Coordinate: 37^o 30’ 13.86” – 1^o 41’ 38,07”
Station LO41 (V. Guadalentín) – Altitude 693m	Coordinate: 37^o 18’ 72” – 1^o 49’ 6,62”

Table 2. Information collected every hour for each station.

Weather station code	Mean wind direction (^{o}) (meWD)	Cooling units (Hf_R)
Min. relative humidity (%) (miRH)	Mean relative humidity (%) (meRH)	Max. relative humidity (%) (maRH)
Sunlight (S)	Rainfall (mm) (R)	Accumulated radiation (W/m^2) (AR)
Mean radiation (W/m^2) (meR)	Max. radiation (W/m^2) (maR)	Mean wind run (km in to hour) (WR)
Min. temperature (^{o}C) (miT)	Mean temperature (^{o}C) (meT)	Max. temperature (^{o}C) (maT)
Mean wind speed (m/s) (meWS)	Max. wind speed (maWS) (m/s)	

4.2. Datasets, evaluation and analysis of experiments

For this initial study, all measured attributes are used, where for attributes with several values for the same measurement (min, med, max) we will use the mean value attribute. The domain of these attributes is shown in Table 3 (the nomenclature of each attribute corresponds to the information reflected in Table 2).

Therefore, the dataset has 104544 instances with 10 attributes joint to the station code. The CMEANS algorithm is used by varying the number of clusters c. CMEANS is executed using the following parameters: $\varepsilon = 0.001$ and c taking values from 1 to 15. Evaluation and analysis of the clusters is performed using the following:

Table 3. Dataset information

Instances number		104544		Per station 8712			Attributes number		10 + 1 (station)		
Attributes information		meWD	Hf_R	meRH	S	R	AR	meR	WR	meT	meWS
Numeric	Minimum	0	-1	5.57	–	0	0	0	0	-9.75	0
	Maximum	360	1	99.98	–	33	4.22	1173.17	1.76	40.45	11.72
Nominal		S	values	{yes,no}							

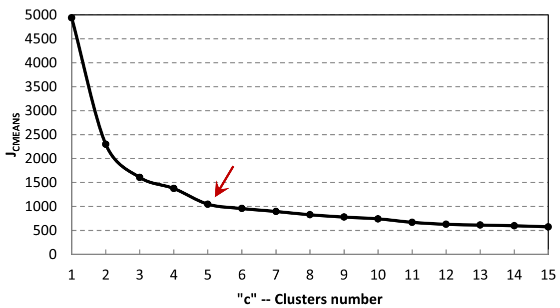
- Elbow Method: The behavior of the function J_{CMEANS} is evaluated and analyzed. A point (or zone of points) is located, the so-called elbow, where the improvement of the function J_{CMEANS} begins to stabilize.
- The attribute of the stations. To analyse each group of the selected cluster we will use the labels of the stations in order to check similar behaviours with respect to different cultivation areas.

4.3. *Results of the experiments*

Executing the CMEANS algorithm for the different values of c, the following values of the function J_{CMEANS} are obtained together with the graph of its behavior (Table 4).

Table 4. Behavior of the function J_{CMEANS} according to the parameter c

c	J_{CMEANS}	c	J_{CMEANS}
1	4939.71	9	779.59
2	1860.06	10	742.77
3	1361,52	11	670.17
4	1183.58	12	628.44
5	1049.29	13	612.78
6	958.19	14	598.47
7	896.99	15	576.11
8	828.01		



In the figure of Table 4 the size of the cluster $c = 5$ is highlighted, which indicates the area where the function J_{CMEANS} is stabilized. The size in instances of the groups of this cluster is the following (in parenthesis the percentage that represents with regard to the total of the dataset): Group 1 = 16029 (14.62%), Group 2 = 16978 (16.24%), Group 3 = 16699 (15.97%), Group 4 = 17471 (16.71%) and Group 5 = 37367 (35.74%).

4.4. *Analysis and discussion of the experiments*

Table 5 shows the characteristics of the groups according to the values of the attributes.

To analyze the results, the domains of the numerical attributes meRH, R, AR, meR, WR, meT and meWS are partitioned in 5 equal parts to discretize them in 5 labels; the domain of the wind direction attribute (meWD) is partitioned using the cardinal points; and the attribute Hf_R is partitioned with three labels to collect the information that indicates how the temperature is inducing in the rest of the annual cycle of many plants (in particular, the fruit trees). Table 6 displays the different centroids of the cluster with 5 groups by means of the labels of the different centroids.

Table 5. Centroids of the 5 groups of the cluster

Group	meWD [0,360]	Hf_R [-1,1]	meRH [5.57,99.98]	S {Y,N}	R [0,33]	AR [0,4.22]	meR [0,1173.17]	WR [0,1.76]	meT [-9.75,40.45]	meWS [0,11.72]
1	253.44	0.20	48.79	N	0.019	0.13	15.10	0.38	12.16	2.52
2	205.12	-0.50	36.95	Y	5.0E-4	1.80	508.83	0.39	17.33	2.59
3	157.86	0.69	75.82	N	0.06	1.78	7.88	0.24	6.60	1.64
4	196.20	0.50	59.70	Y	0.01	1.32	323.46	0.29	10.16	1.93
5	221.01	0.71	77.65	N	0.08	0.02	5.73	0.13	6.05	0.85

Table 6. Centroids with labels: VS-very small, S-small, M-medium, L-large, VL-very large; for MeWD, N∈[337.5^o,22.5^o], NE∈[22.5^o,67.5^o], E∈[67.5^o,112.5^o], SE∈[112.5^o,157.5^o], S∈[157.5^o,202.5^o], SO ∈[202.5^o,247.5^o], O∈[247.5^o,292.5^o], NO∈[292.5^o,337.5^o]; for Hf_R, P∈[0.4,1], C∈[-0.2,0.4], N∈[-1,-0.2]

Group	meWD	Hf_R	meRH	S	R	AR	meR	WR	meT	meWS
1	O	C	M	No	VS	VS	VS	S	M	S
2	SO	N	S	Yes	VS	M	M	S	M	S
3	S	P	L	No	VS	M	VS	VS	S	VS
4	S	P	M	Yes	VS	S	S	VS	S	VS
5	SO	P	L	No	VS	VS	VS	VS	S	VS

In addition, the information collected from the different groups with respect to the 12 stations is the following:

a. Group 5 is the most important with 35.74% of the total instances, and includes the following stations, where the percentage of the instances included of each station is indicated: CR12 (53.55%), JU71 (43.81%), JU81 (48.46%), MO22 (34.62%), CI42 (46.15%), CI52 (51.86%), CR32 (53.76%), AL41 (46.44%) and TP91 (50.25%).

b. Group 3 (15.97% of instances). This group includes the stations: JU42 (64.12%), LO21 (62.90%) and LO41 (64.66%).

c. Group 1 (14.62% of instances). This group includes the stations: CR12 (13.54%), JU71 (23.27%), JU81 (19.43%), MO22 (32.70%), CI42 (20.57%), CI52 (16.06%), CR32 (13.57%), AL41 (20.26%) and TP91 (16.05%).

d. Group 2 (16.24% of instances) contains all stations: CR12 (9.77%), JU42 (10.38%), JU71 (16.30%), LO21 (16.37%), JU81 (15.99%), MO22 (20.44%), CI42 (19.33%), CI52 (17.94%), CR32 (15.53%), AL41 (20.35%), LO41 (12.88%) and TP91 (19.61%).

e. Group 4 (16.71% of instances) also contains all stations with the following weights: CR12 (23.14%), JU42 (22.59%), JU71 (16.62%), LO21 (17.14%), JU81 (16.12%), MO22 (12.24%), CI42 (13.95%), CI52 (14.14%), CR32 (17.14%), AL41 (12.95%), LO41 (20.43%) and TP91 (14.10%).

Analyzing Table 6 and the information related to the stations in each group, the following patterns are observed:

- Note that all groups obtain a very low average rainfall value for the period analysed.
- Groups 1, 3 and 5 have the common characteristics of very low solar radiation and predominantly non-sunny days.

- Groups 1 and 5 include 9 stations located in the Northwest, Central and South Altiplano, Vega del Segura and Cartagena areas. Group 5 differs from Group 1 by showing south-west winds, positive units of cold, while in Group 1 they are null, more humidity, lower temperatures and almost no wind. This indicates that there are periods of time, between the months of December and March, that the areas have two different meteorological behaviours.
- Group 3 includes the rest of the areas, that is, the 3 stations located in the south and south-east Valle del Guadalentín, and the northern Altiplano. This group has similar meteorological conditions to Group 5 showing differences in wind direction and a higher accumulated radiation.
- Groups 2 and 4 have the common characteristic of predominantly sunny days. The information from these two groups indicates that there are two time periods in which all regions show two different behaviours. In one of them, the temperature is low, positive units of cold, almost no wind from the south, low radiation and average relative humidity; while the other has a medium temperature and radiation, low wind speed from the south-west, low humidity, and negative units of cold.

5. A forecast of future objectives

As it has been commented, in this work there are several elements that can be taken into account to improve the process carried out. A preliminary analysis of these elements is presented below:

1. Expand the number of instances, both complementing the periods of the year (spring, etc.) and with new weather stations in the Region of Murcia.
2. Analyze the different clusters in a more adequate way and taking into account different time periods (which allows us to obtain time groups with different behaviors), it is necessary to design the dataset as time series.
3. As shown in Table 2, there are several measurements that have not been used (minimum and maximum values). In the present work, the attributes with average values have been used. In future situations, attributes represented by fuzzy values (both numerical and nominal) could be used in order to collect all the available information on the measurements with the values (min, med, max) and (med, max).
4. Having datasets that can contain fuzzy attributes, the design and implementation of a clustering technique that is capable of supporting and modeling from instances represented with fuzzy attributes is required.
5. Propose a decision support system that reflects these advances and helps the farmer in making decisions regarding his different crops.

6. Conclusions

In this work, a preliminary study on the groupings of geographical areas of the Region of Murcia obtained using a clustering technique is carried out. These groups have been characterized by their weather conditions in the winter season. As a main conclusion we can say that the technique has been useful to find physically separated areas with similar characteristics. This is a promising starting point towards achieving the final goal, that is, helping agricultural sustainability. This will cause to improve the performance of certain agricultural areas based on the best performance of similar areas and the use of crops according to weather conditions.

Acknowledgement

Supported by the projects TIN2017-86885-R (AEI/ FEDER, UE) and RTC-2017-6389-5, granted by the Ministry of Economy and Competitiveness of Spain (including ERDF support).

References

[1] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors. 2018; 18(8):2674.

[2] Jantakat Y. Spatiotemporal Change for Agricultural Distribution from Municipal to Provincial Scales-based Spatial Clustering Analysis. International Journal of Building, Urban, Interior and Landscape Technology. 2019; 14:67-80.

[3] Reddy AD, Dadore B, Watekar A. Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning. Int. J. of Sci. Res. in Science and Technology. 2019; 6(1):485–489.

[4] Villalobos R, Retana J. Un Método para el Pronóstico de Lluvias en Costa Rica: Agrupación de años con características pluviométricas semejantes para la creación de escenarios climáticos. Tópicos Meteoreológicos y Oceanográficos. 2001; 8(1):50–54.

[5] Gavioli A, de Souza EG, Bazzi CL, Schenatto K, Betzek NM. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. Biosystems engineering. 2019; 181:86–102.

[6] MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. In: Le Cam LM, Neyman J, editors. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. 1967; Berkeley: University of California Press, vol. 1 p. 281–297.

[7] Fundación Integra. Informes de "Región de Murcia Digital - Economía - Agricultura". Región de Murcia. Spain. Available from: https://www.regmurcia.com/servlet/s.Sl?sit=c,24,m,3111&r=ReP-23210-DETALLE_REPORTAJESPADRE