Use of Unsupervised Machine Learning Methods for Sugarcane Crop Suitability Evaluation

Roberto F. Silva[1], Alex da S. Sousa[2], Fernando Xavier[1], Emerson Galvani[2], Gustavo M. Mostaço[1], Antonio M. Saraiva[1], Carlos E. Cugnasca[1], Jurandyr L. S. Ross[2]

[1]Department of Computer Engineering and Digital Systems, Polytechnic School – University of São Paulo (USP) Av. Prof. Luciano Gualberto, 380 - Butantã – 05508-010 – São Paulo – SP – Brazil

[2]Geography Department – School of Philosophy, Literature and Human Sciences – University of São Paulo (USP) Av. Prof. Lineu Prestes, 338 - Butantã – 05508-000 – São Paulo – SP – Brazil

{roberto.fray.silva, alex.sousa, fxavier, egalvani, gmostaco, saraiva ,carlos.cugnasca, juraross}@usp.br

Abstract. *Crop suitability evaluation plays an essential role on strategic planning for agricultural activities. Due to future climate change scenarios, there is a possibility that areas previously suitable for certain crops may become inadequate. The rules-based method used to evaluate crop suitability depends on costly field experiments. This paper proposes and evaluates the use of the k-means clustering algorithm for sugarcane crop suitability evaluation in the state of São Paulo, comparing it with the traditional method. The results indicate that it may provide important information for decision making, especially on climate change scenarios and for the suitable and suitable with irrigation categories.*

1. Introduction

Crop suitability estimation is essential for strategic planning and decision making because it allows farmers and the Government to better estimate where a given crop should be planted. It also allows for better irrigation planning and is essential for evaluating future scenarios due to climate change.

For sugarcane crop in the state of São Paulo, it has been done by the São Paulo State Department of Agriculture using rules related to three main variables: temperature, annual water deficit, and water index [CIIAGRO 2008]. Nevertheless, the use of this method (referred to as the traditional method) is costly, in terms of both investment and time, as it demands planting variations of the crops on several locations on the state and evaluating the plants' growth and productivity.

Several important research papers [Junior et al. 2006; Massignam et al. 2017] evaluate crop suitability on different climate change scenarios. Nevertheless, they consider mainly the use of the traditional method. Therefore, they still demand data from productivity in different areas and conditions.

The objective of this paper is to propose and evaluate the use of unsupervised machine learning to improve decision making and providing an alternative source of information for crop suitability estimation for the sugarcane crop in the state of São Paulo, considering different climate change scenarios, and the suitability categories (referred to as zones) currently in use by the traditional method. The data used is related to temperature, rainfall, water deficit and soil type, which are cheap to obtain. We

evaluate clustering and classification metrics and perform a spatial analysis of the implementation of the k-means++ method in comparison to the traditional method.

2. Theoretical foundation

2.1. Sugarcane crop suitability evaluation

For sugarcane cultivation, climate restrictions are obtained using primary data, such as temperature, and secondary data, such as water deficit and water index, which are calculated using other variables such as temperature and precipitation. Table 1 illustrates the sugarcane suitability rules according to a classification developed by the São Paulo State Department of Agriculture [CIIAGRO 2008]:

Table 1. Suitability classification for sugarcane crop

Variables	Rules description
Temperature	- Average annual temperature below 20 ºC: Unsuitable for culture on a commercial scale with maturation problems and frost risks; - Average annual temperature between 20 and 21 ºC: Marginal; - Average annual temperature above 21 ºC: Optimal for the culture.
Annual water deficit	- Annual water deficit of less than 5 mm: Unsuitable; - Annual water deficit between 5 and 10 mm: Marginal; - Annual water deficit greater than 10 and less than 250 mm: Suitable.
Water index	- Annual water index higher than 80: Unsuitable, excess humidity; - Annual water index between 60 and 80: Marginal; - Annual water index below 60 and above -20: Suitable.

Source: Adapted from CIIAGRO (2008).

2.2. Machine learning and k-means clustering

Machine learning is a type of artificial intelligence that gives machines the ability to learn without explicit programming, by discovering patterns from data inputs [Mahdavinejad et al. 2018]. Other machine learning definitions also consider the machine's ability to improve its performance on learning tasks continually.

The machine learning methods can be divided into three main categories: supervised, unsupervised, and reinforcement. Clustering is an unsupervised machine learning method that can be used to identify possible correlations between sets of variables or to group data according to their similarity [Elavarasan et al. 2018]. Among many algorithms for clustering, one of the most used is the k-means, which aims to organize data in a predefined number of clusters. The main objective of the k-means method is to identify groups that have: (i) homogeneous data points inside the group; and (ii) heterogeneous data points between groups. To reach this objective, it uses euclidean distances of the points on the different dimensions [Elavarasan et al. 2018].

The k-means algorithm has been used in many applications, both in agriculture and climatology. Examples of work using this algorithm involve water resources management [Roushangar and Alizadeh 2018], improvements in agricultural production [Huang et al. 2017], crop disease identification [Han et al. 2016], among others. The k-means algorithm was also used to analyze the potential impacts on the habitat of northeastern American tree species [Casajus et al. 2016].

3. Methodology

The methodology used in this paper was composed of five steps, which are:

1. Data gathering and analysis, using the Pandas Python library. The variables used were: latitude, longitude, soil type, average annual temperature in ºC, average monthly temperature in ºC, annual precipitation volume in mm, monthly precipitation volume in mm, water deficit in mm, and temperature in July in ºC. The main data sources collected were: climate data from Brazil [Embrapa 2019], and suitability analysis according to soil types from 261 cities in the state of São Paulo [CIIAGRO 2008];
2. Implementation of the traditional method for sugarcane crop suitability estimation [CIIAGRO 2019], considering four scenarios, based on the research by Junior et al. (2006): (i) current conditions; (ii) IPCC1: increase of 1 ºC in average annual temperature and 15 % in annual precipitation volume; (iii) IPCC2: increase of 3 ºC in average annual temperature and 15 % in annual precipitation volume; and (iv) IPCC3: increase of 5.8 ºC in average annual temperature and 15 % in annual precipitation volume. Five suitability classes were considered, based on [CIIAGRO 2019]: (i) Zone 1 - suitable - optimal for cultivation; (ii) Zone 2 - suitable - low thermal restriction; Zone 3 - marginal - seasonal water deficit area, irrigation required; Zone 4 - marginal - absence of dry period, difficulties in maturation and harvest; and Zone 5 - unsuitable for sugarcane cultivation;
3. Implementation of k-means++ using the scikit-learn library. Several experiments were conducted varying the input variables and the model's hyperparameters;
4. Analysis of the k-means++ implementation considering two categories of quality metrics: (i) traditional classification metrics: as precision, recall, and F1-score; and (ii) supervised clustering metrics: Adjusted Rand score, Mutual info score, Homogeneity score, Completeness score, V-measure, and Fowlkes-Mallows score. Pandas, scikit-learn, and matplotlib libraries were used to calculate and analyze those metrics;
5. Development and analysis of maps considering all four scenarios with the traditional and the k-means++ methods. The ArcGIS software was used to develop maps and the Pandas Python library was used for statistical analysis.

4. Results and Discussions

Table 2 illustrates the results of the traditional classification metrics for the k-means++ implementation in all scenarios analyzed. The 0.000 values represent the fact that the model did not correctly predict any data point in that zone. Values close to 1.000 mean that the model succeeded in predicting more data points in that zone. The F1-score is the most important metric, as it is a harmonic mean of precision and recall. For this metric, values above 0.500 were considered good results, and are highlighted on the table.

For most scenarios, the algorithm performs badly on classifying zones 4 and 5. For zone 2, it performs well in the current conditions and IPCC1. Nevertheless, it is not capable to predict data points that belong to this zone on the other scenarios.

The most relevant results are the considerably good predictions made for zones 1 and 3, especially for the current conditions and IPCC2 scenarios. For example, for the current conditions scenario, prediction for zone 1 presented an F1-score of 0.796, a recall of 0.804 and a precision of 0.787, which can be considered good results for an unsupervised method applied on a small dataset with few features. This indicates that the generated model could be useful for improving decision making between suitable without irrigation (zone 1) or suitable with irrigation (zone 3).

Unlike the traditional classification metrics, which provide results for each category, or cluster, the supervised clustering metrics provide results for the overall

model. In this way, they provide an overall evaluation of the model, indicating if it is a suitable solution for the problem. For all the analyzed metrics, values closer to 1 indicate better results.

Table 2. Results of the k-means++ model for traditional classification metrics

Scenario	Metric	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Current conditions	Precision	0.787	0.627	0.626	0.105	0.000
	Recall	0.804	0.427	0.528	1.000	0.000
	F1-score	**0.796**	**0.508**	**0.573**	0.190	0.000
IPCC1	Precision	0.575	0.496	0.976	0.000	0.000
	Recall	1.000	0.816	0.323	0.000	0.000
	F1-score	**0.730**	**0.617**	0.485	0.000	0.000
IPCC2	Precision	0.602	0.077	0.708	0.000	0.895
	Recall	0.898	0.019	0.810	0.000	0.600
	F1-score	**0.721**	0.030	**0.756**	0.000	**0.718**
IPCC3	Precision	0.474	0.000	0.615	0.000	1.000
	Recall	0.931	0.000	0.533	0.000	0.512
	F1-score	**0.628**	0.000	**0.571**	0.000	**0.677**

Table 3 presents the results for the supervised clustering metrics. For each metric, the highest value was highlighted. The model showed the best results for the IPCC2 scenario, confirming the results observed in the traditional classification metrics. It also presented worse results for the IPCC3 scenario, except for the homogeneity and Fowlkes-Mallows metrics, which showed worse results for the current conditions.

Table 3. Results of the k-means++ model for supervised clustering metrics

Metric	Current conditions	IPCC1	IPCC2	IPCC3
Adjusted Rand score	0.255	0.263	**0.468**	0.231
Mutual info score	0.374	0.443	**0.578**	0.268
Homogeneity score	0.474	0.476	**0.615**	0.548
Completeness score	0.388	0.457	**0.587**	0.281
V-measure	0.427	0.466	**0.601**	0.372
Fowlkes-Mallows score	0.483	0.510	**0.603**	0.577

Figure 1 illustrates the maps of the different scenarios and methods. A spatial analysis of those maps indicates that, as observed with the evaluated metrics, zone 1 (green) and zone 3 (orange) presented the highest similarity between both methods, for all scenarios. This indicates that the k-means++ method was more accurate in predicting those zones. Zone 5, on the other hand, was the one to present the worst results for the k-means++ model. Other important insights were : (i) traditional rules penalized the increase in temperature more than the k-means++ method; and (ii) for the extreme scenarios, k-means++ model increased its performance on the worst class (zone 5) and decrease its performance on zone 2.

Observation (i) is expected since the k-means++ method considers only the distance between the data points on the n-dimensions describing that specific data point. Increasing the penalty on the model did not affect significantly these results, indicating that a higher number of clusters (or zones) could improve the model's results. As for the second observation, more research is needed, to better understand its cause.

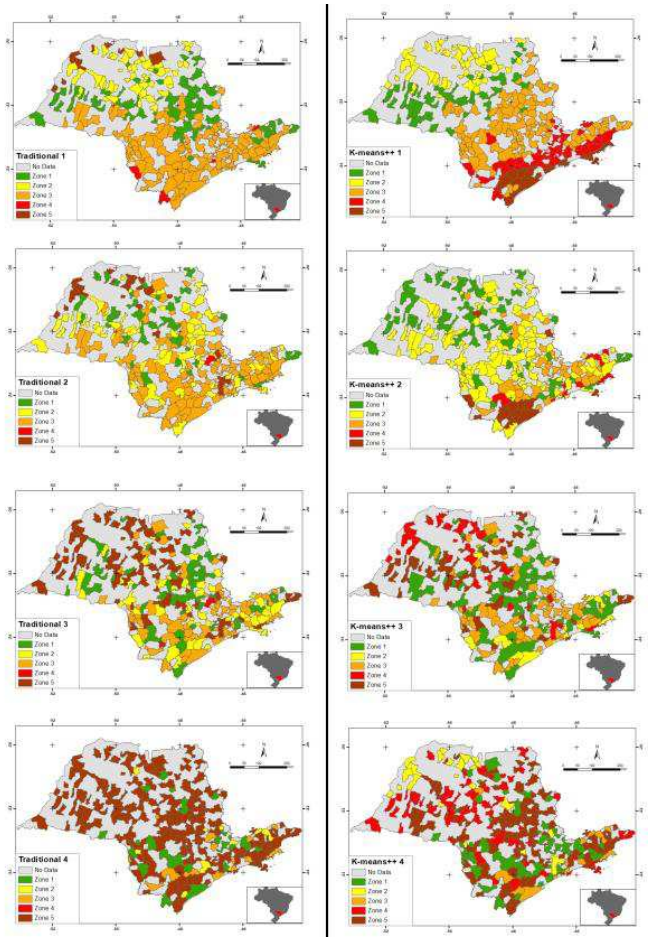

Figure 1. Maps of the classification results for each scenario, for the traditional (left side of the figure) and k-means++ implementations (right side of the figure).

5. Conclusions

Crop suitability evaluation is essential for strategic decision making for farmers and the Government. A rules-based method is traditionally used, based on costly experiments. We analyzed the use of the k-means algorithm as an alternative for estimating crop suitability for sugarcane in the state of São Paulo, concluding that: (i) it presents good overall results for predicting the zone 1 (suitable) and zone 3 (suitable with irrigation) categories; (ii) it presents bad overall results for predicting the zone 4 (marginal with absence of dry period) and zone 5 (unsuitable); and (iii) it presents different behaviors on the different scenarios, with the best results obtained on the IPCC2 scenario.

Therefore, we believe this method is a good alternative for improving decision-making. Further work is related to: (i) evaluating the correct number of clusters based on the structure present in the data; (ii) incorporating more features; (iii) evaluating the

suitability for other crops; and (iv) evaluating the use of supervised learning models. The main limitations observed were: (i) the lack of clustering implementations for crop suitability; and (ii) the lack of open data to incorporate additional features on the model.

Acknowledgments

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, Itaú Unibanco S.A. through the Itaú Scholarship Program, at the Centro de Ciência de Dados (C2D), Universidade de São Paulo, Brazil, and also by the National Council for Scientific and Technological Development (CNPq).

References

Casajus, N., Périé, C., Logan, T., et al. (25 mar 2016). An Objective Approach to Select Climate Scenarios when Projecting Species Distribution under Climate Change. *PLOS ONE*, v. 11, n. 3, p. e0152495.

CIIAGRO (2008). Zoneamento de Culturas Bioenergéticas no Estado de São Paulo - Aptidão Edafoclimática da Cultura da Cana-de-Açúcar. . http://www.ciiagro.sp.gov.br/zoneamento/2008/Zoneamento2008a.htm.

CIIAGRO (2019). Aptidão Edafoclimática da Cultura da Cana de Açúcar. http://www.ciiagro.sp.gov.br/zoneamento/2008/Zoneamento2008a.htm, [accessed on Jul 10].

Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y. and Srinivasan, K. (dec 2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and Electronics in Agriculture*, v. 155, p. 257–282.

Embrapa (2019). Banco de Dados Climáticos do Brasil. https://www.cnpm.embrapa.br/projetos/bdclima/balanco/index/index.html.

Han, L., Haleem, M. S. and Taylor, M. (2016). Automatic Detection and Severity Assessment of Crop Diseases Using Image Pattern Recognition. p. 283–300.

Huang, J., Islam, A. R. M. T., Zhang, F. and Hu, Z. (15 oct 2017). Spatiotemporal analysis the precipitation extremes affecting rice yield in Jiangsu province, southeast China. *International Journal of Biometeorology*, v. 61, n. 10, p. 1863–1872.

Junior, J. Z., Pinto, H. S. and Assad, E. D. (2006). Impact assessment study of climate change on agricultural zoning. *Meteorological Applications*, v. 13, n. S1, p. 69–80.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., et al. (aug 2018). Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, v. 4, n. 3, p. 161–175.

Massignam, A. M., Pandolfo, C., Santi, A., Caramori, P. H. and Vicari, M. B. (2017). Impact of climate change on climatic zoning of common bean in the South of Brazil. *Embrapa Trigo-Artigo em periódico indexado (ALICE)*.

Roushangar, K. and Alizadeh, F. (15 jul 2018). A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. *Journal of Mountain Science*, v. 15, n. 7, p. 1481–1497.