DOI: 10.5748/20CONTECSI/COM/AGB/7257

eLocator: e207257

ASSESSMENT OF SUGARCANE PRODUCTION USING REGRESSION MODELS AND RGB VEGETATION INDICES DATA

Inacio Henrique Yano – https://orcid.org/0000-0003-2698-6309
Embrapa Agricultura Digitalfatec Santana De Parnaíba

Mariana Lopes De Carvalho – https://orcid.org/0000-0002-8119-2264
Faculdade De Tecnologia De Piracicaba (Fatec)

Luis Fernando Sanglade Marchiori
Universidade De São Paulo (Usp) - Escola Superior De Agricultura Luiz De Queiroz (Esalq)

Fabio Cesar Da Silva – https://orcid.org/0000-0003-3733-4592
Embrapa Agricultura Digitalfaculdade De Tecnologia De Piracicaba (Fatec)

ASSESSMENT OF SUGARCANE PRODUCTION USING REGRESSION MODELS AND RGB VEGETATION INDICES DATA

Abstract

Productivity estimates play a crucial role for sugarcane producers and sugar mills in planning production, aligning it with demand forecasts. Manual estimations demand considerable effort and time, prompting exploration into alternative productivity estimation methods such as aerial imaging using drones. Within imaging techniques, productivity estimation occurs indirectly through the analysis of vegetation indices. The widely recognized vegetation index, NDVI, necessitates costly near-infrared (NIR) cameras, making it inaccessible to many producers. Our approach utilized drone imagery captured by more affordable RGB cameras, which are feasible for a larger number of producers. We applied six regression models alongside a stacking model that amalgamated these six models for estimating sugarcane production using the eight RGB vegetation indices. Initial tests revealed a Mean Absolute Percentage Error (MAPE) of less than 13%. This level of accuracy is considered favorable when benchmarked against similar studies and presents encouraging prospects for future research.

Keywords:drone; imagery; machine learning; uav;yield.

1 Introduction

Productivity estimates play a critical role in assisting sugarcane producers and sugar mills to plan their sugar, ethanol, and related product output. These estimates are pivotal for adjusting production according to demand forecasts (Mawandha et al., 2022). Furthermore, they aid in forecasting profitability, influencing investment projections in sugarcane fields, harvesting, transportation processes on farms, and industrial plant operations within the sugar and alcohol sector. Therefore, these estimates are indispensable for professional management in this industry.

There are various methods for estimating productivity. Manual estimation involves using a mobile application, offering benefits such as backup options and utilizing artificial intelligence to assist in recording plant counts per hectare (Mawandha et al., 2022). Satellite imagery is another method, especially useful for covering vast areas. However, this approach has limitations due to the Normalized Difference Vegetation Index (NDVI) association with biomass, leading to challenges with low spatial and temporal resolution (Singla et al., 2015).

In contrast to labor-intensive manual data collection and the spatial-temporal limitations of satellite imagery, unmanned aerial vehicles (UAVs) offer an alternative for yield estimation through imagery. UAV imagery provides superior spatial resolution, and the imaging process can be repeated as needed, offering flexibility and accuracy (Poudyal et al., 2022).

Yield estimation often relies on vegetation indices such as NDVI, which effectively indicate plant health. However, calculating NDVI requires the use of Near-Infrared (NIR) wavelengths (Stamford et al., 2023). These wavelengths are accessible only through multispectral and hyperspectral cameras, which tend to be costly and out of reach for many producers. Consequently, several studies have turned to RGB vegetation indices as alternatives to gauge plant health and biomass production (Bakacsy et al., 2023; Sanches et al., 2018). RGB cameras are more affordable compared to multispectral ones, making them accessible to a larger number of farmers. Additionally, they are lighter, enabling longer drone flight times (Yano et al., 2017).

The aim of this study is to assess and choose machine learning regression models and ensemble techniques for estimating sugarcane yield using eight RGB vegetation indices. The resulting models will serve as effective tools for sugarcane producers, empowering them to make informed decisions, particularly regarding cash flow management and investment planning.

2 Methodology

As outlined in the preceding section, this study aimed to generate models for estimating sugarcane production using RGB vegetation indices. To accomplish this, we applied six regression models that utilized a combination of eight RGB vegetation indices to determine the most effective model. This section delineates the imaging environment, vegetation index computation for dataset creation, and the generation of regression models for estimating sugarcane production.

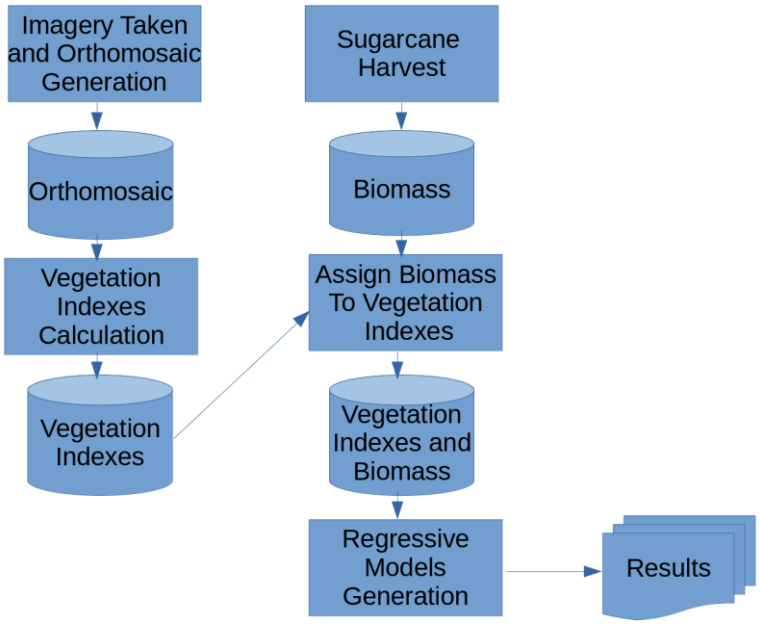
Figure 1 illustrates the flowchart depicting the model generation process. Initially, images were captured from the field using a drone, and subsequently, an orthomosaic was generated. The AgroAzul Company facilitated image capture and orthomosaic generation for this study.

From the received orthomosaic, crops within the experimental field plots were delineated and utilized for computing the vegetation indices.

Sugarcane harvesting occurred some days after the drone imagery. During this process, the weights of each plot within the experimental field were recorded in a spreadsheet (Table 1). These weights signify the biomass produced in the experimental field and were matched with the previously calculated vegetation indices to establish the dataset reflecting field production for each plot.

For generating the regression models, six algorithms were employed using the field production dataset, resulting in seven regression models for predicting sugarcane production. Each algorithm produced one model, and an additional stacking model, an ensemble combining the six models, aimed to improve individual model outcomes (Abdallah et al., 2022). The selected metric for result comparison was the Mean Absolute Percentage Error (MAPE), as it offers a more understandable outcome, facilitating comparison between predicted and real values to determine the magnitude of differences. This metric was preferred over Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) due to its interpretability in assessing error comparisons with original (real) values.

Figure 1 – Regressive models generation flowchart

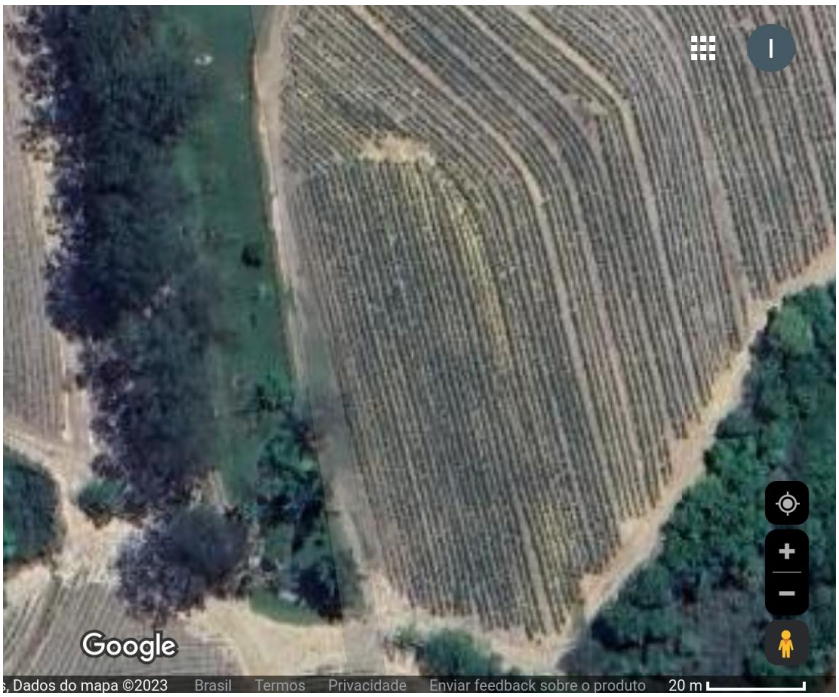


Source: Authors

2.1 Environment

The experiment site was on a farm located at -22.682012, -47.837749 in the municipality of Piracicaba (Figure 2). The experimental field has 40 plots (Figure 3). Each plot is a rectangle with a length of 10 m and width of 7.5 m, with an area of 75 m^2. Sugarcane was planted throughout the experimental field. At the time of harvest, the production of each plot was harvested and weighed separately to construct Table 1. Before the harvest, the AgroAzul Company takes the RGB Imagery and generated an orthomosaic of the experimental field.

Figure 2 – Experimental Field



Source: Authors

Figure 3 – Experimental field divided into 40 plots

Source: Authors

Table 1 – Plot number, weigh in kg and weigh in Ton per hectare

Plot	Weight (kg)	Ton per Hectare
1	790,06	105,34
2	823,71	109,83
3	814,86	108,65
4	752,86	100,38
5	713,89	95,18
6	635,94	84,79
7	637,71	85,03
8	797,14	106,29
9	742,23	98,96
10	953,03	127,07
11	940,63	125,42
12	882,17	117,62
13	958,34	127,78
14	1006,17	134,16
15	823,71	109,83
16	940,63	125,42
17	901,66	120,22
18	921,14	122,82
19	869,77	115,97
20	846,74	112,90
21	898,11	119,75
22	850,29	113,37
23	733,37	97,78
24	747,54	99,67
25	930,00	124,00
26	914,06	121,87
27	956,57	127,54
28	782,97	104,40
29	772,34	102,98
30	738,69	98,49
31	901,66	120,22
32	956,57	127,54
33	682,00	90,93
34	802,46	106,99
35	779,43	103,92
36	806,00	107,47
37	657,20	87,63
38	988,46	131,79
39	637,71	85,03
40	481,83	64,24

Source: Authors

2.2 Vegetation indices Calculation

A Python program calculated the eight RGB Vegetation indices used in this work. The formula of each vegetation index is listed below and uses the letters R, G, and B, which represent the colors red, green, and blue, respectively:

a) Excess of Green (ExG) described by Woebbecke et al. (1995):

$$ExG = 2G - R - B \quad (1)$$

b) Excess of Red (ExR) proposed by Meyer et al. (1998):

$$ExR = 1{,}4R - B \quad (2)$$

c) Excess Green minus Excess Red (ExG−ExR) (Meyer & Neto, 2008):

$$ExG–ExR = ExG - ExR \quad (3)$$

d) Normalized Difference Vegetation Index (NDI) (Perez et al., 2000):

$$NDI = (G - R)/(G + R) \quad (4)$$

e) Visible Atmospherically Resistant Index (VARI) (Cen et al., 2019):

$$VARI = (G - R)/(G + R - B) \quad (5)$$

f) Green Leaf Index (GLI) (Eng et al., 2019):

$$GLI = (2G - R - B)/(2G + R + B) \quad (6)$$

g) Red Green Blue Vegetation Index (RGBVI) (Bendig et al., 2015):

$$RGBVI = (GG - RB)/(GG + RB) \quad (7)$$

h) Green Red Ratio Vegetation Index (GRRI) (Lu et al., 2021):

$$GRRI = G/R \quad (8)$$

2.3 Dataset Creation

After the harvest of the sugarcane crop, each plot of the experimental field had its biomass measured. The dataset for the regressive models has the plot identification, and the biomass weights joined with the vegetation index calculated for the sub-images extracted from the orthomosaic. The eight vegetation indices are ExG, ExR, ExG-ExR, NDI, VARI, GLI, RGBVI, and GRRI.

2.4 Models Generation

For Models Generation, a specific Python program using Sklearn Library applies six regressive algorithms, which makes the stacking ensemble generate seven models, where the results can be seen in a boxplot graph for result comparison. This program also applies the cross-validation technique because it makes the full use of the data to provide more precise error estimates (Bergmeir et al., 2014). In the cross-validation, the training and validation sets of data are never the same, preventing overfitting (Berrar, 2019). The algorithms and their hyperparameters are listed in Table 2 below.

Table 2 – List of algorithms used for sugarcane production estimation

Algorithm	Acronym	Hyperparameters
KNeighborsRegressor	knn	n_neighbors=8
DecisionTreeRegressor	cart	max_depth= 8,max_features= 'auto', min_samples_leaf= 10, min_samples_split=0.01, criterion='absolute_error'
SVR	svm	kernel='rbf',C=0.1, epsilon=0.1
RandomForestRegressor	rf	criterion='absolute_error',max_depth=8, max_features='auto', min_samples_leaf=10,min_samples_split=0.01, n_estimators=200
HistGradientBoostingRegressor	xgb	learning_rate =0.1, max_iter=100, max_depth=2, min_samples_leaf=20, max_leaf_nodes=25
LinearRegression	lr	
Stacking	stacking	

Source: Authors

3 Results

The results were divided into two parts. The first part is the linear regression metrics of the eight vegetation indices, just for comparison and to know which one had a better correlation to the biomass measured in the harvest procedure. The second part is the result of the regression using the six regression algorithms and the result of the stacking ensemble to select the better model for sugarcane production estimation.

3.1 Linear Regression metrics of the eight vegetation indices

The linear regression metrics of the eight vegetation indices are listed in Table 3.

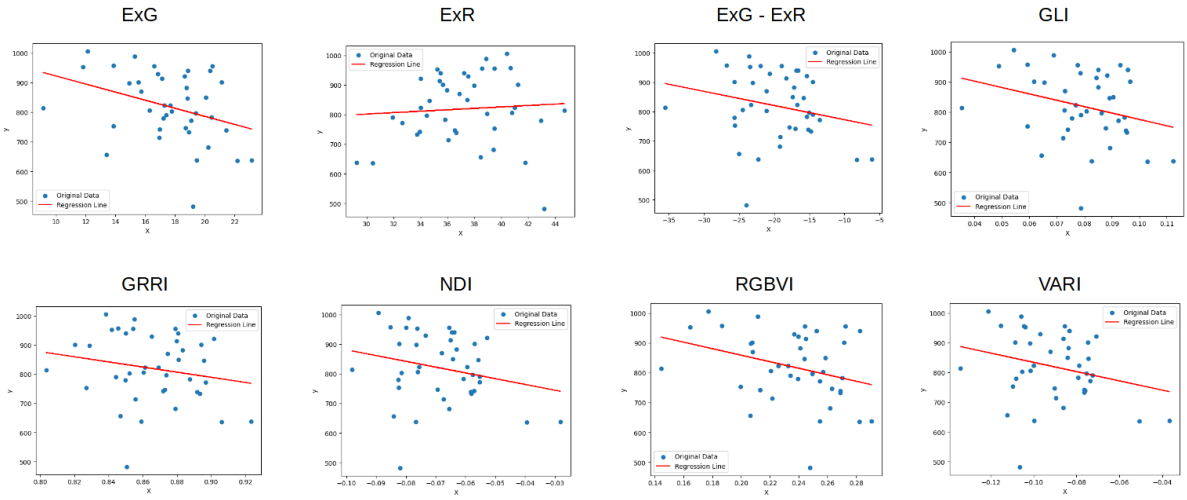
Table 3 – Linear regression metrics of the eight vegetation indices

Metric	ExG	ExR	ExG-ExR	NDI	VARI	GLI	RGBVI	GRRI
MAPE	0.1146	0.1233	0.1203	0.1204	0.1195	0.1176	0.1154	0.1219
MAE	87.65	93.76	91.48	91.46	90.84	89.70	88.08	92.91
MAE TCH	11.69	12.5	12.20	12.19	12.11	11.96	11.74	12.39

Source: Authors

Observing Table 3, the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE) indicate that the difference between the original value measured and the predicted value was around 12%. The Excess of Green presented the best results, followed closely by RGBVI. Figure 4 shows the graphs of the eight vegetation indices.

Figure 4 – Linear regression graph of the eight vegetation indices



Source: Authors

3.2 Apply six regression algorithms and stacking ensemble to vegetation index dataset

The results of six regression algorithms and stacking ensemble applied to the eight vegetation indices are listed in Table 4.

Table 4 – Metrics of the six regression algorithms and stacking ensemble applied to the eight vegetation indices

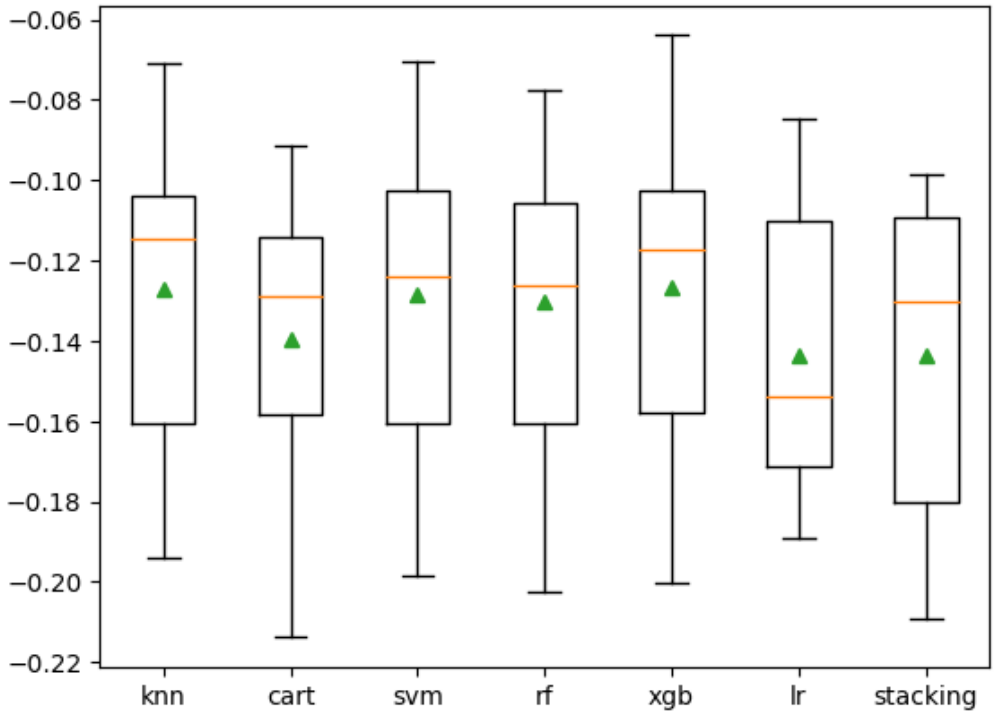
Metric	knn	cart	svm	rf	xgb	lr	stacking
MAPE	0.1273	0.1396	0.1285	0.1302	0.1268	0.1438	0.1437
MAE	95.05	105,79	97.98	98.04	96.67	109.82	106,55
MAE TCH	12.67	14.11	13.06	13.07	12.89	14.64	14.21

Source: Authors

Table 4 shows that XGBoost (XGB) was the best algorithm for a MAPE, followed by K Nearest Neighbor(kNN). For MAE, kNN achieved the best results, followed by XGB. The Support Vector Machine (SVM) algorithm takes the third position in these two metrics. The MAE values in ton per hectare (TCH) for kNN, XGB, SVM, and Random Forest (RF) are better than those presented by Caetano (2017). For this dataset, the stacking algorithm did not work well, taking the second worst position better only the linear regression algorithm. The kNN algorithm achieved these results by using seven neighbors. The XGB had almost the

same results, even changing the value of the hyperparameters. XGB and kNN overcame the other regressive models.

Figure 5 – Boxplot of the six regression algorithms and stacking ensemble applied to the eight vegetation indices



Source: Authors

4 Conclusion

This work presented regressive models that predict sugarcane production based on RGB Imagery taken from drones. This model's prediction can estimate sugarcane biomass production with less than 13% of errors. Unfortunately, the stacking ensemble did not work well for the dataset generated in this experiment. But, the XGB regressive model, which is a kind of ensemble, had the best results together with the kNN regressive model.

In future works, we intend to construct a larger dataset to improve the results because the kNN regressive model had better results with seven neighbors. We also start capturing images two or three months before the harvesting to enforce the prediction aids in sugarcane commercialization.

5 References

Abdallah, E. B., Grati, R., & Boukadi, K. (2022, June). A machine learning-based approach for smart agriculture via stacking-based ensemble learning and feature selection methods. In 2022 18th International Conference on Intelligent Environments (IE) (pp. 1-8). IEEE.

Bakacsy, L., Tobak, Z., van Leeuwen, B., Szilassi, P., Biró, C., & Szatmári, J. (2023). Drone-Based Identification and Monitoring of Two Invasive Alien Plant Species in Open Sand Grasslands by Six RGB Vegetation Indices. Drones, 7(3), 207.

Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., ... & Bareth, G. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. International Journal of Applied Earth Observation and Geoinformation, 39, 79-87.

Berrar, D. (2019). Cross-Validation.

Bergmeir, C., Costantini, M., & Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. Computational Statistics & Data Analysis, 76, 132-143.

Caetano, J. M., & Casaroli, D. (2017). Sugarcane yield estimation for climatic conditions in the state of Goiás. Revista Ceres, 64, 298-306.

Cen, H., Wan, L., Zhu, J., Li, Y., Li, X., Zhu, Y., ... & He, Y. (2019). Dynamic monitoring of biomass of rice under different nitrogen treatments using a lightweight UAV with dual image-frame snapshot cameras. Plant Methods, 15, 1-16.

Eng, L. S., Ismail, R., Hashim, W., & Baharum, A. (2019). The use of VARI, GLI, and VIgreen formulas in detecting vegetation in aerial images. Int. J. Technol, 10, 1385-1394.

Lu, J., Cheng, D., Geng, C., Zhang, Z., Xiang, Y., & Hu, T. (2021). Combining plant height, canopy coverage and vegetation index from UAV-based RGB images to estimate leaf nitrogen concentration of summer maize. Biosystems Engineering, 202, 42-54.

Meyer, G. E., Hindman, T. W., & Laksmi, K. (1999, January). Machine vision detection parameters for plant species identification. In Precision agriculture and biological quality (Vol. 3543, pp. 327-335). SPIE.

Meyer, G. E., & Neto, J. C. (2008). Verification of color vegetation indices for automated crop imaging applications. Computers and electronics in agriculture, 63(2), 282-293.

Perez, A. J., Lopez, F., Benlloch, J. V., & Christensen, S. (2000). Colour and shape analysis techniques for weed detection in cereal fields. Computers and electronics in agriculture, 25(3), 197-212.

Poudyal, C., Costa, L. F., Sandhu, H., Ampatzidis, Y., Odero, D. C., Arbelo, O. C., & Cherry, R. H. (2022). Sugarcane yield prediction and genotype selection using unmanned aerial vehicle-based hyperspectral imaging and machine learning. Agronomy Journal, 114(4), 2320-2333.

Sanches, G. M., Duft, D. G., Kölln, O. T., Luciano, A. C. D. S., De Castro, S. G. Q., Okuno, F. M., & Franco, H. C. J. (2018). The potential for RGB images obtained using unmanned aerial vehicle to assess and predict yield in sugarcane fields. International journal of remote sensing, 39(15-16), 5402-5414.

Singla, S. K., Dubey, O. P., & Garg, R. D. (2015). Application of geoinformatics in automated crop inventory. International Journal of Computer Applications, 975, 8887.

Stamford, J. D., Vialet-Chabrand, S., Cameron, I., & Lawson, T. (2023). Development of an accurate low cost NDVI imaging system for assessing plant health. Plant Methods, 19(1), 9.

Sumesh, K. C., Ninsawat, S., & Som-ard, J. (2021). Integration of RGB-based vegetation index, crop surface model and object-based image analysis approach for sugarcane yield estimation using unmanned aerial vehicle. Computers and Electronics in Agriculture, 180, 105903.

Woebbecke, D. M., Meyer, G. E., Von Bargen, K., & Mortensen, D. A. (1995). Color indices for weed identification under various soil, residue, and lighting conditions. Transactions of the ASAE, 38(1), 259-269.

Yano, I. H., Mesa, N. F. O., Santiago, W. E., Aguiar, R. H., & Teruel, B. (2017). Weed identification in sugarcane plantation through images taken from remotely piloted aircraft (RPA) and KNN classifier. J. Food Nutr. Sci, 5(6), 211.

6 Acknowledgments

The authors thank the Embrapa-Embracal project 30.21.90.004.00.00 - Improvement of technical recommendations for correcting soil acidity and its phytotechnical implications in sugarcane for the support provided to carry out this work. The authors also thank to AgroAzul Company for the excellent work in imagery services provided.