

Remote Sensing and Machine Learning Modeling to Support the Identification of Sugarcane Crops

CARLOS LOZANO-GARZON¹, (Senior Member, IEEE), GERMÁN BRAVO-CÓRDOBA¹, HAROLD CASTRO¹, (Senior Member, IEEE), GEOVANNY GONZÁLEZ-RODRÍGUEZ¹, DAVID NIÑO¹, HAYDEMAR NUÑEZ¹, CAROLINA PARDO¹, AURELIO VIVAS¹, YUBER CASTRO², JAZMIN MEDINA², LUIS CARLOS MOTTA², JULIO RENE ROJAS², AND LUIS IGNACIO SUÁREZ²

¹Systems and Computing Engineering Department, Universidad de los Andes, Bogota 111711, Colombia

²Banco Agrario de Colombia, Bogota 111711, Colombia

Corresponding author: Carlos Lozano-Garzon (calozanog@uniandes.edu.co)

ABSTRACT One of the main concerns of agricultural financing institutions is to make sure the loans they grant are used for the stated objective when the loan was requested. Specifically, when Banco Agrario de Colombia grants loans for crop farmers, it schedules verification visits to the cultivation sites to check if the crop stipulated in the loan agreement exists and assess its health. These visits are challenging to make due to the number of visits over vast areas that they need to schedule, lack of trained personnel, and difficulty of access. This article proposes a software tool, based on a machine learning model for processing free satellite imagery, to support the bank's identification of non-compliant crops with the investment plan before making field visits, minimizing the loss of investment by focusing on those areas to prioritize the visits. Sugarcane along the department of Boyacá, Colombia was chosen as the case of study. Free access satellite imagery through the Colombian Data Cube (CDCol) was used and machine learning models were applied on them to classify the land and predict the presence of the crop, a Random Forest model achieved an overall F1-score of 91% using Landsat-8 imagery and a K-nearest Neighbors model achieved an overall F1-score of 98% using Sentinel-2 imagery.

INDEX TERMS Investment monitoring, Landsat-8, machine learning, remote sensing, Sentinel-2, sugarcane crops.

I. INTRODUCTION

Banco Agrario de Colombia (BAC) is the Colombian government entity in charge of implementing financial support initiatives for the country's farmers. One of BAC's main objectives is to support Colombian farmers through loans that allow them to grow and harvest agricultural products.

Agricultural and rural credits are regulated by Colombian state legislation and, therefore, their "processing and granting must comply with the provisions contained in Colombian Laws 16 of 1990 and 1731 of 2014, as well as the other regulations that add or modify them. They must also comply with the Resolutions issued by the National Agricultural Credit Commission, the Circulars of the Financial Superintendence of Colombia, Superfinanciera, or the Superintendency of the Sector of Cooperatives, Supersolidaria. In addition, the

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

agricultural sector development financing agency, Finagro, has a Service Manual that states: "The credits are granted by financial intermediaries, entities that have a direct relationship with the beneficiary, which must monitor the correct use of monetary resources, and certify to Finagro compliance with the regulations that govern them" [1].

In January 2018, Finagro modified the Title Five of its Service Manual, adjusting its "Commitments, monitoring, control and verification procedure" of those operations registered with Finagro, "which directly affects the verification process of the investment controls and the commitments that the beneficiaries of the credits with re-discount resources and those who receive subsidies from the National Government must assume" [2]. The new modification established that "the client must make the expenses and investments contemplated in the financed project within the foreseen time and report any changes, such as: financed item, investment impact, either due to climatic or phytosanitary problems,

among others, that may partially or totally affect the investments, in order for the BAC to evaluate the viability of authorizing the proposed modification and present it again to Finagro” [1].

Currently, BAC carries out the process called Controls of Agricultural Investment just after the end of the credit-granting process. The aim of this process is the follow-up and control of the investments made by the bank’s customers in order to detect, in time, any inconvenience in the productive development of the investments, for timely decision making.

The monitoring process involves the following activities: (1) Generating the list of visits according to the coverage area established by the bank, (2) visits assignment to the staff of the bank according to the capacity and coverage established for each of them, (3) the assignments are checked to ensure that they correspond to the coverage area, and the feasibility of the visit, (4) contact the clients in order to verify the status of the investments and scheduling the follow-up visit, (5) carrying out the visits that involve staff travel to planned sites, (6) collecting information about the status of the crop, georeference points of the farm where the crop is located, and photographic evidence of the vegetative state of the premises; and, after the visit, (7) validating the investment status and (8) making decisions about the investment progress [3].

The great demand for productive agricultural projects financed by the bank represents a major problem that currently makes it impossible to effectively monitor projects advancements which financing comes from the BAC. Currently, the BAC has one hundred and one (101) specialized advisers in charge of monitoring around 880,000 productive projects approved per year. They barely manage to cover 36,000 visits, which falls short of the visits required by law - a minimum sample of 10% of the approved projects - that must be reported to Finagro.

This imbalance makes it difficult to cover 100% of the investment monitoring. According to projections made by the Investment Controls and Appraisals Office of BAC, the mentioned imbalance could be corrected increasing the human resource capacity; this hiring would have represented an annual payroll expense for the BAC of more than \$1M USD [1].

Concerning the visits, into a report provided by the Agricultural Technical Monitoring Sub-management, the five (5) main causes of non-compliance were: (1) The farmer did not carry out the investment, (2) client not found, (3) diversion of resources, (4) public order, and (5) climatic factors.

Another issue in the Colombian countryside is the poor condition of the tertiary road network, which connects municipal capitals and small towns or towns with each other; this network represents 69% of the national road network. This problem, joined with the public order situation in Colombia, hinders the work of the Bank’s commercial advisers in placing loans on site and in monitoring the agricultural investment requested by control entities for agricultural loans, which operate under Finagro conditions.

Last but not least, according to the National Agricultural Census conducted in 2014 [4], 70% of the food produced in the country comes from small producers who carry out agricultural production work on their farms, most of which are not less than 5 hectares in size. Therefore, the crops areas are not extensive.

To deal with these issues, an application to support the agricultural investment control process is suggested. Specifically, we propose the use of free satellite (Landsat-8 and Sentinel-2) images through the Open Data Cube (ODC) infrastructure, to handle the images storage and processing, and, based on this, develop a Machine Learning model to predict the presence of specific crops in areas of interest of the bank.

In consequence, a tool is provided to determine whether loans given to farmers to plant a specific crop are actually being used to fulfill the loan’s purpose, by verifying the geospatial location of the property and identifying the crop. The aim is to support the identification process of crops with non-compliance in the investment plan, validate areas with fraud problems before making field visits, minimizing the loss of investment by focusing on the areas to be visited, and prioritizing those visits that must be reported to Finagro.

Taking into account the diversity of crops financed by the bank and Colombia’s territorial extension, the focus of this case study is the identification of sugarcane crops in the Department of Boyacá (around 23,000 km^2).

According to reports presented by Agronet, sugarcane is one of the crops with the greatest economic and social importance for Colombia, due to the high number of people who work in the sugarcane life cycle and its high per capita consumption. In the same way, in a report on sectoral indicators for the period between 2008 and 2013, published by the Ministry of Agriculture and Rural Development in 2013, the sugarcane cultivation is the second in the country in generation of direct and indirect jobs, after coffee, with a contribution of 11.5% [5].

Likewise, 2018’s report from the Ministry of agriculture [6] remarks the importance of sugarcane to the country’s economy; furthermore, it showed that more than 350,000 families develop this crop. Also, it generates about 287,000 direct jobs, equivalent to 45 million wages per year and employs 12% of the economically active rural population. The departments with the greatest productive influence in this subsector are: Boyacá, Cundinamarca, Cauca, Antioquia, Santander, Nariño, Valle del Cauca, Tolima, Caldas, Norte de Santander, Risaralda and Huila, where 83% of the cultivated area is concentrated.

Figures provided by Fedepanela - Fondo de Fomento, for the year 2017 the country reached a total of 228,976 hectares planted, a harvested area of 205,156 hectares, an average yield of 5.66 tons of panela (which is one of the final products made from the sugarcane juice that, by successive boiling, loses moisture and solidifies into blocks) per hectare and a total production of 1,284,141 tons of panela, in 29 out of 32 departments of the country covering 564 municipalities.

The planted areas were centered in the departments of Cundinamarca, Antioquia, Santander, Boyacá, Cauca, Nariño, Tolima, which represent 77.35% of the total sugarcane production; it should be noted that only 4 departments, Cundinamarca, Antioquia, Santander and Boyacá have 53.37% of the area planted nationwide [5]. Likewise, in [7], it was evidenced that in terms of figures related to production, the Department of Boyacá occupied the third place out of 29 in the country's total production with 13.7%.

This representative place is still kept within BAC because according to data provided by the Investment Controls and Appraisals department of BAC, 21% out of the 1,217 monitoring and control visits to the investment in 2019 were carried out in the Department of Boyacá (251 visits). Similarly, BAC reports that, for sugarcane, 24,282 credits were granted through Finagro between 2018 and 2019, for a value of 286,130 million COP and of this figure, 22,189 credits were rediscounted, for a value of 101,920 million COP; which confirms its importance. Furthermore, sugarcane is the fourth (4th) agricultural activity with the highest number of credits approved by the bank after coffee, fruit trees and plantain.

This paper is organized as follows: section II describes the state of the art in crop monitoring. In section III, we explain the methodology used for the development of the identification models of sugarcane in the Department of Boyacá, emphasizing the use of data cubes for the construction of training data and highlighting the fact that, although the information sources are different (Landsat-8 and Sentinel-2), the methodology for developing the models is the same. The general discussion and insights of the process are presented in section IV and, finally, section V presents the conclusions and future work.

II. LITERATURE REVIEW

Studies on crop classification date back to 1987, when Landsat Multispectral Scanner System (MSS) and Thematic Mapper (TM) images were used to apply maximum likelihood (probabilistic) algorithms, visual interpretation, unsupervised classification, and threshold-based segmentation. Results from studies made between 1987 and 1997 achieved precision values between 70% and 93%, where the best precision was achieved using the active satellite ERS-1 with the maximum likelihood algorithm for rice grading [8]. Regarding the use of multispectral imaging, in 2015, a Chilean research group analyzed the use of Landsat-8 images for the phenological classification of fruit tree crops [9]. In this study, they compared the performance of three classifiers applied to the images (linear discriminant analysis (LDA), Random forests (RF), and Support Vectorial Machine (SVM)) using different operations on images such as NDVI, normalized difference water index (NDWI), and time series using all image bands. As a result, they found that using time series with all image bands provides a more accurate classification than using NDVI and NDWI, specifically applying LDA and time series over reflections in each band. The same year, Kharat and Musande used the k-means algorithm to map

cotton crops using Landsat-8 images achieving an accuracy of 98.01% for a k-value (number of groups in the algorithm) of 10 [10].

Concerning delimiting sugarcane crops, Wang *et al.* (2020) [11] proposes the joint use of optical multispectral images, obtained by Landsat-8 and Sentinel-2 satellites, and SAR images, obtained by the Sentinel-1 satellites, to generate annual maps of sugarcane at the field scale over large regions. Through the use of geo-referenced polygons, the authors obtain the base pixels to calculate spectral indices (NDVI, EVI, LSWI, and mNDWI); subsequently, they proposed the use of a pixel-phenological algorithm, supported by time series and classification trees, to determine the presence of sugarcane in a given region. After performing the system test, they obtained an overall identification accuracy of 96%. The main challenges reported in the study were: (i) the small size of sugarcane crops in this province (< 1 ha); (ii) the presence of other surrounding crops, such as rice or corn; (iii) the topography of the region; and (iv) the frequent cloud cover.

Shendryk, Davy & Thorburn (2020) conducted a study to predict field-level sugarcane yield in the northeast Queensland region of Australia. In this study, they used Sentinel-1 and Sentinel-2 satellite imagery in combination with climate, soil and elevation data. Authors implemented four different types of predictive machine learning models (Random Forest, Gradient Boosting, Extreme Trees and Extreme Gradient Boosting) in order to forecast the cane yield (t/ha), commercial cane sugar (CCS, %), sugar yield (t/ha), crop varieties and ratoon numbers. The model with the best performance was Gradient Boosting, using this model they found that sugarcane varieties could be mapped with an accuracy of up to 73.4%, while the differentiation of planted and ratoon crops exhibited the lowest accuracy of 45.4%. The main challenges reported in the study were: (i) the climate variability in the region; (ii) soil types; and (iii) harvesting processes in the area.

Concerning the delimitation of the sugarcane crops in Colombia, the Cane Research Institute, Cenicaña, published in 2009 "Principles and Applications of Remote Sensing in Sugarcane Crops in Colombia" [12]. This book constitutes a guide for sugarcane remote sensing using different statistical methods. First, it discusses the importance of spectral vegetation indices as it generates an efficient estimation of soil vegetation cover. In second place, statistical methods are proposed aiming to detect sugarcane. These methods are: Principal components analysis, linear analysis of spectral mixtures, Tasseled cap transformation (index), and texture treatment in the image. Physical methods, genetic algorithms, and hybrid methods are also mentioned (hybrid methods include decision trees, support vector machines, and neural networks). The research that led to the publication of the mentioned book referred to studies that used Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat-5/7, National Oceanic and Atmospheric Administration (NOAA) and 'Satellite pour l'Observation de la Terre' (SPOT-4/5) sensor images.

Bastidas, E. *et al.* [13] evaluated the applicability of MODIS data to predict the amount of harvest in Colombia; this publication concluded that linear models in combination with vegetation indices such as EVI had an accuracy of 74% to estimate final production at an early stage (from the fifth month of cultivation). Murillo, P. *et al.* [14] analyzed a methodology for monitoring sugarcane in the Cauca River Valley, also using satellite images from the MODIS platform; it was found that it is possible to monitor cultivated areas larger than 6.25 ha with moderate resolutions (250 meters to 1000 meters). This study used a combination of regressions with the EVI vegetation index. Other research performed in Colombia by Murillo, S. *et al.* [15] used images from the Landsat 7 ETM+ satellite to detect and discriminate sugarcane varieties in Valle del Cauca; the method used was a combination of vegetation indices such as NDVI, RVI, leaf area index (LAI), atmospheric resistant vegetation index (ARVI) and the adjusted soil vegetation index (SAVI) in addition to a supervised classification using the maximum likelihood algorithm, which assumed that the bands had a normal distribution. A principal component analysis (PCA) was also performed and revealed that the best indices were GNDVI (green difference normalized vegetation index) and GVI (green vegetation index). An accuracy of 80.8% was achieved for the period between 4 and 5 months on large crop areas.

Based on the results of this and other studies, it can be concluded that accurate crop mapping is possible using satellite images [16]. However, despite progress, there are recurrent challenges in the use of remote sensing for the purpose of monitoring small crops with satellite remote sensing. With the use of supervised Machine Learning models, getting training data, in this case images labeled with polygons of the crops of interest, is a constant challenge, since it is a costly and time-consuming process [17]. Additionally, the culture methodology, the region where it is being cultivated, and the temperature, among other factors, entail a variability of the characteristics among crops [18], which can cause different reflectance values. Higher spatial and temporal resolution can positively impact some of the challenges; however, the combined use of multiple image sources also brings a challenge to align the different bands at different resolutions.

To conclude, although the results with sugarcane are promising, the conditions for panela sugarcane are different and should be analyzed independently (region, crop area, and varieties). However, with the review carried out of the most relevant studies in Colombia around satellite remote sensing, there are cross-cutting challenges; the following should be highlighted:

- 1) Variable reflectivity due to factors such as moisture, leaf pigments, physiological status, and morphological characteristics of the species [19].
- 2) Changes in soil reflectance; this can occur due to tides (in coastal areas), rain, and, in general, water on the leaves, which produces a fall in the reflectance of the red band and near infrared compared to dry soils [20].

- 3) Lack of standardization [WG] that can lead to duplication of efforts and increased expenditure of resources.

III. SOLUTION PROPOSAL

From the detailed analysis of the sugarcane cultivation and a deep understanding of its climatic, morphological and contextual factors, the methodology that allows us to obtain sugarcane crop identification models using both Landsat-8 and Sentinel-2 is explained in Figure 1. This methodology is based on the Machine Learning life cycle that includes 5 stages which are implemented as follows: (1) Satellite Data Acquisition, (2) Ground Data acquisition, (3) Data preparation, (4) Model training, and (5) Model evaluation. These steps are elaborated on the subsequent sections.

A. SATELLITE DATA ACQUISITION

Satellite images from the Landsat-8 and Sentinel-2 sensors were acquired from the USGS (United States Geological Survey) and the ESA (European Space Agency) Copernicus Open Access Hub platforms respectively. These platforms provide analysis ready surface reflectance (SR) products; accordingly, Landsat-8 SR images were obtained from the Collection 1 - Level 2 repository and the Sentinel-2 SR images were obtained from the L2A repository. The Land Surface Reflectance Code (LaSRC) applied by the USGS platform to Landsat-8 raw images is detailed in [21], [22], also the SR products obtained from this algorithm are explained. On the other side, Sentinel-2 SR products are generated by ESA through the Sentinel-2 Correction algorithm (Sen2Cor) as described in [23]. Table 1 presents the bands and vegetation indices considered for each sensor, vegetation indices are computed as described in the methodology proposed in Section III.

TABLE 1. Features for the training samples in Sentinel-2 and Landsat-8 data sets.

Feature	Sentinel-2	Landsat-8
red	x	x
green	x	x
blue	x	x
nir	x	x
swir1	x	x
swir2	x	x
pixel_qa		x
scl	x	
narrow_nir	x	
water_vapor	x	
veg5	x	
veg6	x	
veg7	x	
ndvi	x	x
evi	x	x
evi2	x	x
rvi	x	x
savi	x	

B. GROUND DATA ACQUISITION

The activities developed in the data acquisition process include the field visits programmed by BAC to the sugarcane

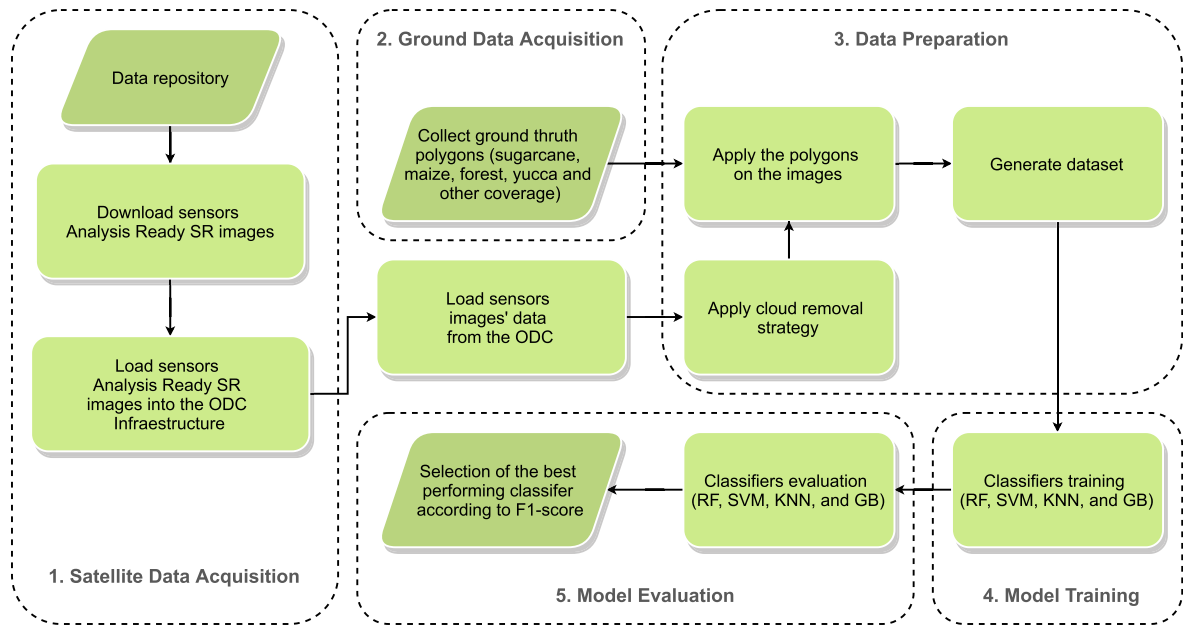


FIGURE 1. Proposed Methodology Depicted as a Workflow.

crops. In these activities, the crops delimitations were geo-referenced which were then turned into polygons expressed in Keyhole Markup Language (KML). Context information about those polygons, including age of the crop, variety, density, and whether this was the only plant contained in the polygon (mixing different crops is a common practice among some farmers) was also requested.

Using the geolocated polygons, Landsat-8 and Sentinel-2 images covering the study area were collected. It is important to note that having the goal of training a multi-class classification algorithm, BAC provided not only sugarcane polygons but also maize, forest, yucca and other coverage. The gathered satellite imagery dates ranged from the day the visit was made back to the month the crop was first planted; this was done in order to increase the sample size and to make sure we included all phenological stages of the crop. All these images were stored in the data cube and studied to understand their characteristics.

Specifically, the bank supplied 40 polygons delimiting the areas of sugarcane crops to be analyzed. However, after the validation process, only 28 polygons were further studied. 12 polygons were discarded since they contained multiple crops (eg, sugarcane and maize). Figure 2 depicts the variety types of sugarcane represented in the set of polygons. The variety RD7511 is the most represented one with 16 polygons of the total set. There are also 12 polygons of other varieties, these varieties are palmireña, common, and ZC.

The age of the polygons integrated in the set are mostly represented between four and seven months with a total of 14 polygons in this range, as shown in Figure 3.

As we mentioned before, the 28 polygons collected on land were exported to KML files; every file was associated with its corresponding metadata located in a csv file. The files

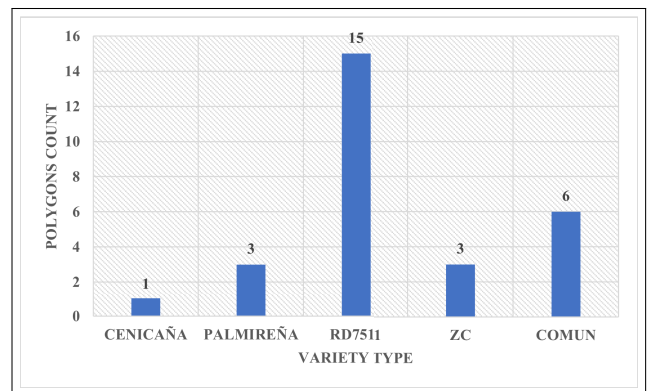


FIGURE 2. Sugarcane polygons seed variety type.

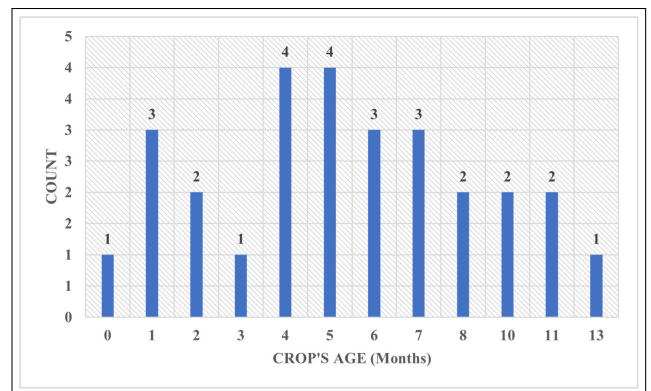


FIGURE 3. Sugarcane polygons crop's age in months.

describe crop's age, variety type, KML file location in the file system, and KML creation date, where the KML creation date tells us the date on which the crop's age in months was registered.

C. DATA PREPARATION

In this phase the data sets needed to build the classification models were generated from the spectral bands and the ground truth data was used to label the pixels. Then, exploratory analysis techniques (statistical measurements like mean, median, standard deviation, outliers and data distributions) were used to know the characteristics of these data sets and to verify their quality. Based on this knowledge, we resampled unrepresented classes, eliminated outliers and standardized the values to improve data representation for learning algorithms.

1) SPECTRAL INFORMATION EXTRACTION ALGORITHM

The 28 KML files along with a csv file containing the metadata of each KML were fed into the spectral information gathering algorithm, with the aim to create the Sentinel-2 and Landsat-8 training data sets. The algorithm carried out the following steps: (1) read a KML file and the metadata associated with it (2) extract the KML file coordinates and KML creation date (3) generate a bounding box of the KML polygon based on its coordinates, (4) query the ODC for an image matching the bounding box and KML creation date, (5) extract spectral information of every point within the polygon boundaries as a vector of features, (6) add the metadata of the polygon to the vector of features, and (7) place the vector data of every collected point in a row of a csv file.

In addition to the training data sets, the algorithm also provides images serving validation purposes. These images depict which points were collected on every satellite image so that we can validate the correctness of the spectral information gathering algorithm and the data sets generated. When images were validated, we noted that low confidence cloud points covering sugarcane polygons were part of the collected Landsat-8 and Sentinel-2 data sets.

2) CLOUDS REMOVAL STRATEGY

Both satellite sensors, Sentinel-2 and Landsat-8, contain quality bands that are useful to determine, in general terms, the type of coverage that the image has at the pixel level.

Accordingly, pixels are classified by these quality bands into several categories including cloud, vegetation, non-vegetation, water, among others in Sentinel-2 but only clear, clouds, water, snow and terrain occlusion pixels in Landsat-8. Dense clouds are correctly classified by these quality bands so, using the cloud mask provided by the quality band, these pixels were removed from the data sets. However, with this approximation we still found low confidence clouds or cirrus pixels in the training data sets that were not detected as such by the quality bands.

In Sentinel-2, sparse clouds or cirrus were being classified as non-vegetated ground as shown in Figure 4. This observation was used to extract a second version of the Sentinel-2 data set, filtering out pixels that were classified as non-vegetation. This was supported by the fact that the sugarcane crop is classified as vegetation from the second

month. With this process, only vegetated pixels within polygons labeled as sugarcane were taken as part of the training data set which yielded to the best results.

Landsat-8 images, unlike Sentinel-2 ones, do not provide pixel quality band classes such as vegetated and non-vegetated ground, that enable cirrus clouds discrimination.

Here, a heuristic was formulated to automate the identification of cirrus and programmatically remove those images of the set to be considered in the spectral information gathering procedure. The blue band provides a leeway in identifying thin clouds. This approach consists in the calculation of the mean for the blue band values of pixels in an image, then replicate this calculation in the image time series and identify particularly bright timesteps.

Figure 5 presents a time series analysis for one of the images considered in the spectral information gathering procedure. We noted that values for the blue mean reflectance higher than 500 reflectance units (ru) represented images with cirrus or low confidence clouds. This approach is applied after pixels classified as clouds by the quality bands are excluded from the data to remove any remaining timesteps that are particularly bright. Finally, images exhibiting a discontinuous behaviour in time were removed. Removing thin clouds also yielded better results for the Landsat-8 data set.

3) DESCRIPTION OF THE GENERATED DATA SETS

The resulting data sets contained 35686 training examples (pixels of satellite images that represented sugarcane crops) in the case of Sentinel-2 imagery, and 1169 pixels in the case of Landsat-8. These resulting numbers correspond to 22.6% and 32% respectively of the initial total number of pixels. This was due to clouds and defective pixels.

Other coverage such as urban zone, water, forest, bare soil, sand, rocks, yucca and maize were also identified and processed, for Sentinel-2 and Landsat-8 images, with the spectral information gathering algorithm, the resulting data sets comprise the coverage shown in Table 2. Furthermore, Table 1 presents the bands and vegetation indices considered for each data set. Since the vegetation indices are calculations over the bands, we decided to add them to the data set as new types of bands in order to have more information, increasing the size of the training data provided to the algorithms.

Figures 6 and 7 describe the Landsat-8 and Sentinel-2 sugarcane final data sets by age respectively. From these figures we can see that crops between one and two months of age are best represented in the Landsat-8 data set and crops between one and six months are best represented in the Sentinel-2 data set.

D. MODEL TRAINING

In this process, multiple classification algorithms such as Random Forests, K-Nearest-Neighbors, Support vector Machine (SVM), Neural Networks and Gradient Boosting were applied. Before applying data pre-processing techniques, data sets were divided into training set and test set (80 % for training and 20 % for testing). The pre-processing

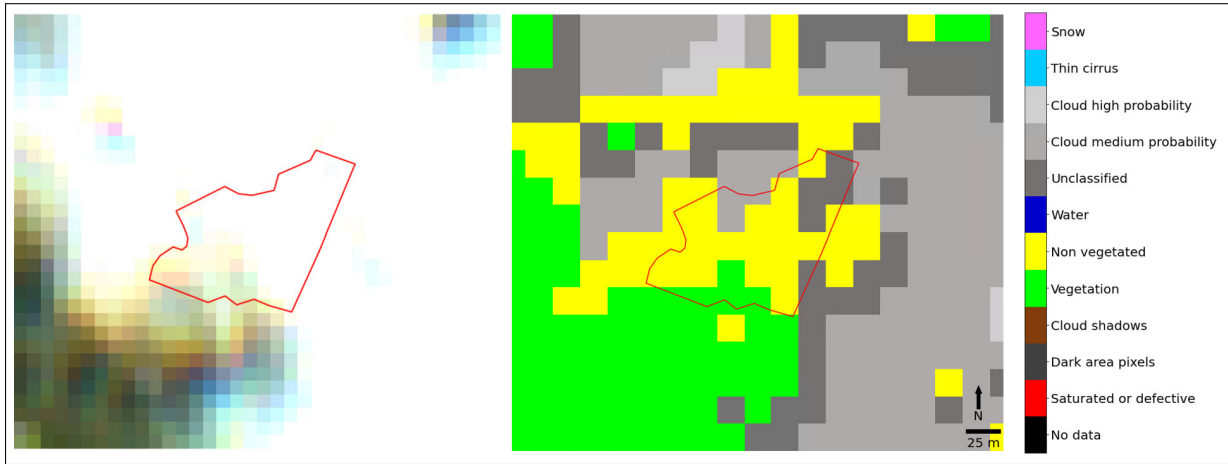


FIGURE 4. Scene classification over sparse cloud in Sentinel-2.

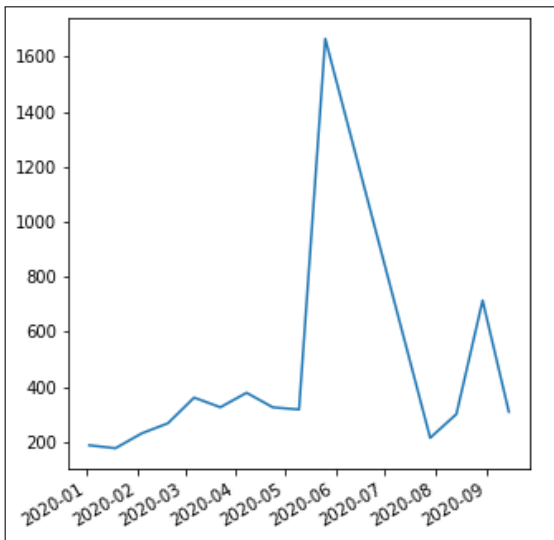


FIGURE 5. Blue mean calculation time series.

TABLE 2. Classes considered in Sentinel-2 and Landsat-8 data sets and number of samples per class.

Coverage	Sentinel-2	Landsat-8
sugarcane	35686	1169
urban zone	188298	4450
water	545018	95301
forest	204299	44547
bare soil	19750	8561
sand	1976	0
rocks	7952	494
yucca	903	0
maize	1149	0
cloud	0	3000
high confidence cloud	0	3000
low confidence cloud	0	3000
medium confidence cloud	0	3000
cloud shadow	0	3000

included balancing the unrepresented classes using resampling. The resampling rate was obtained by applying cross validation; however, experiments were also conducted with

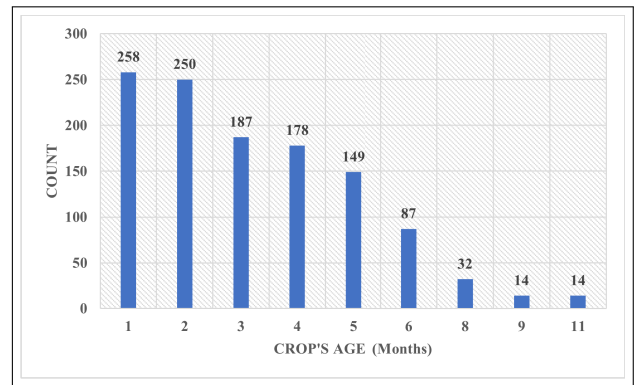


FIGURE 6. Sugarcane Landsat-8 data set crop's age count.

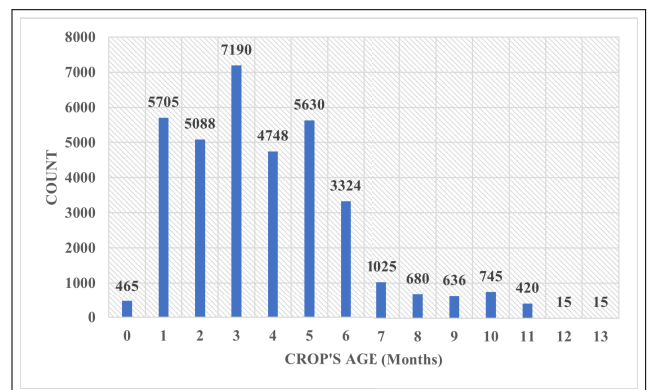


FIGURE 7. Sugarcane Sentinel-2 data set crop's age count.

imbalanced data to determine their effect on the performance of the algorithm. In addition, to determine how vegetation indexes influenced the model's ability to identify sugarcane we also used data sets without that excluding such indexes.

As we mentioned before, five learning algorithms were used for the construction of the classification models. To calibrate these algorithms a search for the best values of

hyperparameters was made; for this purpose, k-fold cross validation technique was applied to training set using $k = 10$. Once the hyperparameter values were obtained, a model was built based on them and applied to the test set to determine its generalization performance on new data. The mentioned hyperparameters include for Landsat-8 and Sentinel-2 models are detailed in Table 3.

TABLE 3. hyperparameter configuration of the models.

	Landsat-8	Sentinel-2
Random Forest	Criterion: Entropy Max Depth: 50 Number of Estimators:100	Class weight: Balanced Criterion: Entropy Max Depth: 25 Min Samples in Leafs: 5 Number of Estimators:120
SVM	C:100 Gamma: scale Decision function shape: one versus one	C: 100 Gamma: scale Decision function shape: one versus rest
KNN	Number of Neighbours: 7 p: 1 Weights: Distance	Number of Neighbours: 1 p: 1 Weights: Uniform
Gradient Boosting	Learning Rate: 0.05 Max Depth: 5 Max Features: Auto Number of estimators: 120	Learning Rate: 0.1 Max Depth: 5 Max Features: Auto Number of estimators: 120

E. MODEL EVALUATION

The models were evaluated on the test set using standard classifier metrics. Based on this analysis, the best model was selected and tested on new polygons provided later by BAC.

1) PERFORMANCE METRICS

Performance of the classifiers obtained was measured using the well-known recall, precision and F1-score metrics; Recall measures positive accuracy, indicating how many examples of this class are correctly classified (is also known as the True Positive rate or Sensitivity); Precision measures how many examples qualified as positive actually belong to this class; and F1-score provides the geometric mean of these two measurements.

$$\begin{aligned}
 \text{Recall} &= \frac{\text{True positives}}{(\text{True positives} + \text{False negatives})} \\
 \text{Precision} &= \frac{\text{True positives}}{(\text{True positives} + \text{False positive})} \\
 \text{F1} &= \frac{(2 * \text{recall} * \text{precision})}{(\text{recall} + \text{precision})}
 \end{aligned}$$

2) EVALUATION OF CLASSIFICATION PERFORMANCE

In this section, we present the results obtained for the classifiers (Random Forests, SVM, Nearest Neighbors, and Gradient Boosting) generated from the Landsat-8 and Sentinel-2 training data sets along their variations; unbalanced without vegetation indices, and balanced with vegetation indices. The classification performance is evaluated on the corresponding test data sets. The average values for recall, precision, and

f1-score are shown for different validations of the classifiers on the test sets.

a: ABOUT LANDSAT-8 CLASSIFIERS

Recall, precision and F1-score metrics for the Landsat-8 classifiers are depicted in Tables 4 and 5. The first table describes the performance of the classifiers that were generated from the unbalanced data set; the second describes the performance of the classifiers generated from the balanced counterpart of the data set.

TABLE 4. Results from Landsat-8 models, Unbalanced data set.

Model	Landsat 8- Unbalanced data set					
	With vegetation indexes			Without vegetation indexes		
	Recall	Precision	F1-score	Recall	Precision	F1-score
KNN	0.83	0.84	0.83	0.84	0.82	0.83
RandomForest	0.90	0.92	0.91	0.90	0.92	0.91
SVM	0.69	0.75	0.71	0.70	0.76	0.71
GradientBoosting	0.91	0.91	0.91	0.90	0.90	0.90

TABLE 5. Results from Landsat-8 models, Balanced data set.

Model	Landsat 8 - Balanced data set					
	With vegetation indexes			Without vegetation indexes		
	Recall	Precision	F1-score	Recall	Precision	F1-score
KNN	0.83	0.76	0.79	0.84	0.79	0.81
RandomForest	0.91	0.85	0.87	0.91	0.85	0.86
SVM	0.80	0.72	0.74	0.82	0.73	0.76
GradientBoosting	0.92	0.86	0.89	0.92	0.87	0.89

As shown in Table 4 Random Forest algorithm achieved the best overall F1-score classification performance, 91% trained with and without vegetation indexes. However, the one classifier generated with the imbalanced-without-indices data set was the best sugarcane classifier, since it delivered 72% F1-score for sugarcane classification against 70%. In terms of individual classes, 9 out of 11 classes achieved a F1-score higher than 84% in both classifiers. The confusion matrix of the best sugarcane classifier is shown in Table 6.

To conclude, we found that the use of vegetation indices such as; NDVI, EVI, EVI2 and RVI did not improve the sugarcane classification accuracy. Although it was expected that the use of the NDVI would improve the classification accuracy as reported on related reports [24], the combination of this index with others caused a negative incidence in the classification. As a result of this observation, the use and evaluation of alternative combination of vegetation indices is proposed as future work. At the same time, considering the amount of sugarcane data that we managed to obtain from Landsat-8, 1416 samples, and the sensor resolution per pixel 30 m^2 , data about 4.248 ha was collected which is short in comparison with the sentinel-2 data set and other Landsat-8 sensor related reports [24]. The low representation of the sugarcane class may cause the model to have reduced capacity to generalize to new data, since the spectral variability in the

TABLE 6. Confusion matrix for Random Forest On Landsat-8.

Classified as	Urban Zone	Water	Sugarcane	Bare Soil	Rocks	Forest	cloud	Shadow	low_conf_cl	high_conf_cl	med_conf_cl
Urban Zone	2437	0	18	4	33	12	21	1	0	1	142
Water	0	1804	3	0	0	0	0	0	1	1	0
Sugarcane	8	4	1713	0	0	0	0	0	27	1	0
Bare Soil	1	0	0	57075	1	4	6	12	0	0	17
Rocks	9	0	2	0	500	10	0	54	6	5	100
Forest	2	0	3	3	9	5135	1	0	0	0	30
cloud	37	0	0	28	0	13	207	0	0	0	6
Shadow	1	0	0	3	36	0	0	26721	3	0	17
low_conf_cl	0	2	4	0	0	0	0	0	1807	3	0
high_conf_cl	1	4	12	0	0	0	0	0	20	1749	4
med_conf_cl	296	1	2	49	124	137	8	73	4	4	1122

cane found in crops is large and this variability could not be sufficiently represented in the data set.

Despite these limitations, we consider that the resulting classifier proof a significant performance in cane classification. This model can be improved as more data for sugarcane and other coverage surrounding the crop is available. Also, the use of other vegetation indices and combination of them should be considered for the improvement.

b: ABOUT SENTINEL-2 CLASSIFIERS

As with the Landsat-8 classifiers, recall, precision and F1-score metrics generated for the Sentinel-2 classifiers, for both the unbalanced and balanced data sets were analyzed and are presented in Tables 7 and 8.

TABLE 7. Results from Sentinel-2 models, Unbalanced data set.

Model	Sentinel 2 - Unbalanced data set					
	With vegetation indexes			Without vegetation indexes		
	Recall	Precision	F1-score	Recall	Precision	F1-score
KNN	0.80	0.83	0.81	0.96	0.96	0.96
RandomForest	0.92	0.73	0.78	0.92	0.98	0.94
SVM	0.91	0.89	0.88	0.93	0.71	0.75
GradientBoosting	0.70	0.79	0.73	0.87	0.89	0.87

TABLE 8. Results from Sentinel-2 models, Balanced data set.

Model	Sentinel 2 - Balanced data set					
	With vegetation indexes			Without vegetation indexes		
	Recall	Precision	F1-score	Recall	Precision	F1-score
KNN	0.83	0.76	0.77	0.96	0.89	0.92
RandomForest	0.91	0.73	0.78	0.97	0.81	0.87
SVM	0.86	0.90	0.87	0.94	0.72	0.76
GradientBoosting	0.84	0.87	0.84	0.89	0.90	0.89

The model that achieved the best overall F1-score is the KNN algorithm as shown in table 7. In terms of individual classes, this model achieved over 84% for every class and 7 out of 9 classes achieved an F1-score of 94% or more. Specifically, the F1-score of this model over sugar cane is 98%, the classes that achieved the lowest F1-score were Maize and Yucca with 88% and 85% respectively. However, it is worth noting that these were the least represented classes in the training data set, as shown in Table 2.

A result of classification over a Yucca crop is shown in Figure 8. On the other hand, there were classes that achieved a 100% F1-score on the test data set, these were urban zone, water, forests and bare soil.

As it was the case with the Landsat-8 classifiers, the best Sentinel-2 classifier was trained without the vegetation indexes; Including the vegetation indexes in the training data set not only did not improve the performance of the model, but it negatively impacted its accuracy. Performance of models trained with and without vegetation indexes can be contrasted in tables 7 and 8.

Also, as we can see in tables 7 8, for KNN and Random forests algorithms an unbalanced data set enhances their performance and, in contrast, SVM and gradient boosting algorithms achieve better results when trained with balanced data sets.

Results of classification over other images are shown in Figure 9 and Figure 10. The first figure shows a classification over a sugar cane crop area, and the second figure shows the classification over a cloudy area that contains a maize crop.

It is important to note that the clouds and cloud shadows classified in the second figure are classes that, as mentioned in section III-C2, are contained in the sentinel scene classification band. Table 9 shows the confusion matrix of the types of cover the algorithm was trained for.

IV. DISCUSSION

A. CONCERNING THE USE OF REMOTE SENSING DATA

Intrinsic features in remote sensing such as temporal, spatial, spectral, and radiometric resolution introduce several challenges when considering land cover detection and classification tasks, in particular, crop detection. It is important to note that these tasks strongly depend on the quantity and quality of the information obtained from the scenes of the different remote sensors.

Regarding temporal resolution, Landsat-8 sensor offers up to 2 scenes per month, while Sentinel 2A and 2B sensors provide 6 scenes per month. By increasing the temporal resolution, the number of scenes per month may require more storage capacity. For Sentinel the required capacity is around 6GB (1GB per scene); therefore, obtaining the information of a specific scene from Sentinel-2 for a whole year represents to the users a 72GB storage requirement.

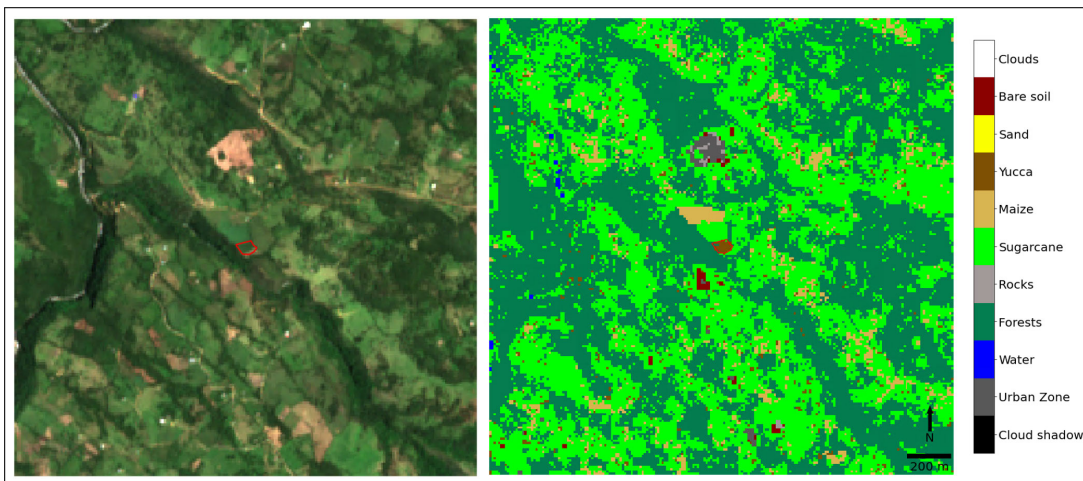


FIGURE 8. Scene classification over yucca polygon in Sentinel-2.

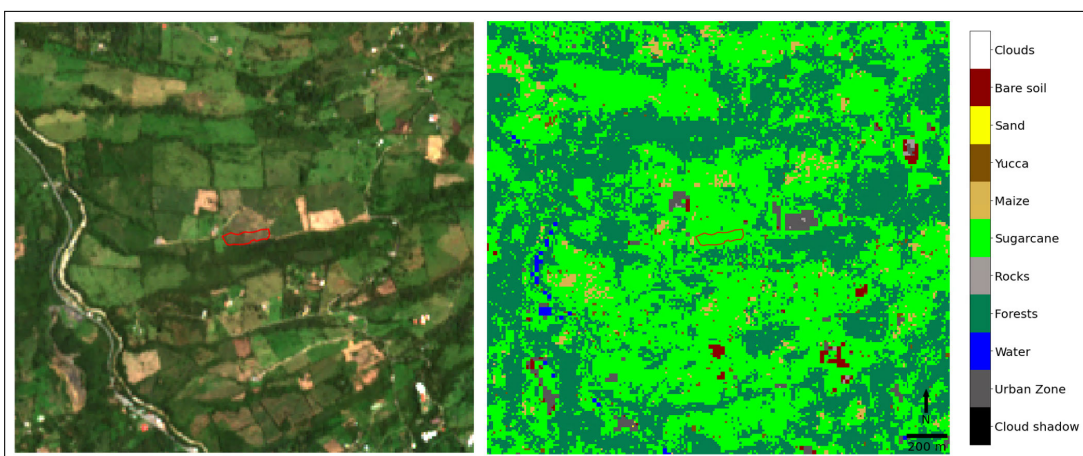


FIGURE 9. Scene classification over cane polygon in Sentinel-2.

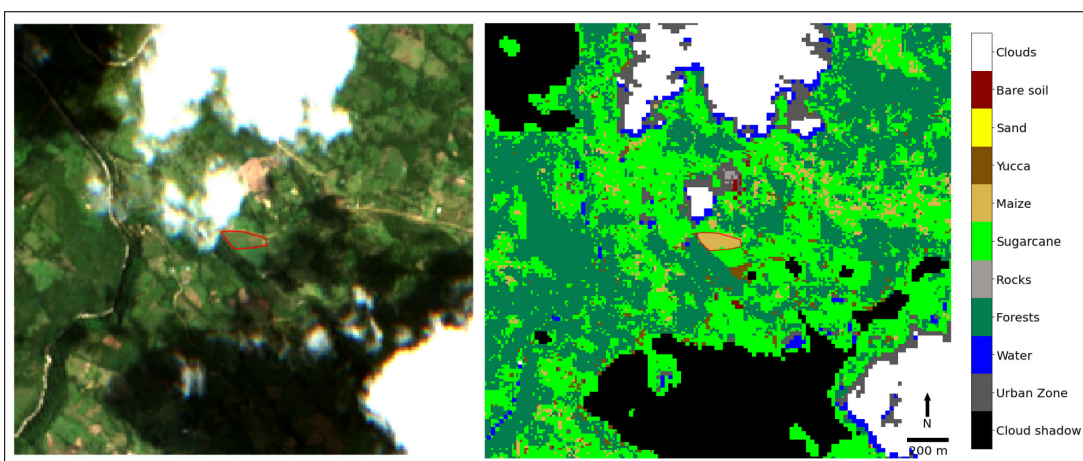


FIGURE 10. Scene classification over cloudy maize polygon in Sentinel-2.

Specifically, for the department of Boyacá, approximately 90% of the territory can be covered using 6 scenes from Sentinel-2, which corresponds to 432GB of storage

per year. The above panorama proposes challenges related to the storage and processing for this increasing data volume.

TABLE 9. Confusion matrix for Nearest Neighbors on Sentinel-2.

Classified as =>	Urban Zone	Water	Forests	Rocks	Sugarcane	Maize	Yucca	Sand	Bare Soil
Urban Zone	37946	0	0	34	0	0	0	49	0
Water	0	108848	0	0	0	0	0	0	0
Forests	0	0	40832	0	44	4	8	0	0
Rocks	23	0	0	1529	0	0	0	0	0
Sugarcane	0	0	92	0	6924	17	23	0	0
Maize	0	0	6	0	27	190	0	0	0
Yucca	0	0	9	0	20	0	166	0	0
Sand	3	0	0	0	0	0	0	402	0
Bare Soil	0	0	0	0	0	0	0	0	3811

Concerning radiometric resolution, slight changes on the crop are difficult to perceive by a sensor.

The atmosphere is composed of gases that cause distortion of the image by the interaction of light with the gases (diffraction). This challenge can be divided into two: the first, related to the distortion of the images due to the gases that make up the atmosphere, even though they allow light to pass through; and the second, related to the appearance of clouds that block the passage of light towards the earth's surface.

To face the first challenge, we developed an algorithms to calculate the impact of this layer of gases, to correct distortions on the satellite image that these gases produce. These algorithms are usually based on the use of a dark surface to determine how an area should look without the atmospheric effects; taking this type of surfaces as a base, the algorithm can predict and counteract the effect of gases on the image.

Secondly, the appearance of clouds in images avoids the correct detection of the ground. To mitigate this inconvenience, algorithms are used to detect their presence and, in this way, only those pixels that have a low probability of clouds are used. Also, radar images like Sentinel-1, provided by active sensors, help to avoid this kind of problems.

B. CONCERNING THE USE OF MACHINE LEARNING ALGORITHMS

The comparative analysis carried out with four learning algorithms and different data sets revealed that the best algorithms were Random Forest and KNN. From these results we can conclude that it is possible to use machine learning techniques to build models that allow the identification of sugarcane crops in the Boyacá region, using data from free access satellite images (Landsat-8 and Sentinel-2).

However, in order to build the labeled data sets needed to apply the modeling techniques, BAC had to reprocess information the had already gathered and include new steps in their visits that were new to their staff. It is important then to establish mechanisms that facilitate the generation of data sets from the moment a loan is granted.

C. CONCERNING THE USE OF THE OPEN DATA CUBE

The use of the python-based API ODC allowed a fast analysis of the remote sensing information. Using this API enabled users to request the pixels that were interesting for analysis directly, instead of manually individual satellite files. Before using the ODC, merging different bands for spectral analysis required a manual resampling method due to the different

resolutions of bands coming from different remote sensing sensors. In contrast, requesting information of different bands with the ODC automatically resamples the bands into a desired resolution and returns them into a single variable ready for analysis.

A process was created in Jupyter notebooks for analysing areas of interest which had multiple options for requesting the information to the ODC. One of the most used ones was requesting pixels by polygon, which returned a square surrounding the desired polygon. This is a highly replicable process, and analysts can change the desired polygons, dates and bands by only specifying them in a set of variables, new analysts can change the variable values to classify a new area.

D. CONCERNING THE FUTURE USE OF GENERATED MODELS FOR CROP MONITORING IN THE CONTROLS OF AGRICULTURAL INVESTMENT PROCESS AT BAC

Although to the problems identified, satellite images constitute a valuable source of information on land surface data. For instance, they would allow with great agility and precision, the geospatial location of the properties presented as a guarantee of credits, as well as the identification of the crop developed in the mentioned property.

In this way, this solution facilitates to the area of Control and Appraisals of BAC, directly responsible for the monitoring and control process, the verification of effective compliance with the conditions agreed in the loan origination stage. Such verification would allow to identify deviations from the investment plan established for the crop, making a filter to identify crops with non-compliance in the investment plan, validating areas with fraud problems, before making a field visit, thus minimizing the loss of the investment by focusing on the areas to visit, prioritizing field visits to those that will be reported to Finagro.

Additionally, it allows the monitoring of crops of products sown with resources disbursed by the Bank in a specific area, by recognizing the area and identifying anomalies in the crops that are the object of investment. This can be done at a property level but also at a regional and even national levels, optimizing the use of resources and the establishment of informed policies within the Bank.

V. CONCLUSION AND FUTURE WORK

This paper presents the development of a software tool, based on a machine learning model for processing free satellite

imagery, with the aim to support the BAC in the Controls of Agricultural Investment process in order to identify crops with non-compliance in the investment plan before making field visits and thus prioritize those visits. As a case study, we selected the identification of “panela” sugar cane crops since it is one of the most important economic and social crops for Colombia.

Based on the results obtained of this work, we found that it is possible to generate reliable models that identify “panela” sugarcane crops in the Boyacá region from free access satellite images. These results reinforce the aim of the BAC to continue the exploration of remote sensing imagery in order to identify the characteristics of production projects supporting the investing control process. However, the generated models are susceptible to many improvements. Some of them are:

- To improve the acquisition of field information, through the capture of crop lots from origination and a protocol more focused on getting useful information for training the Machine Learning models.
- To improve the schedule of visits, in order to get better-balanced information collected training the model. This includes the age of the crops, the variety of cane grown, and whether there are combinations of crops, among others.
- Generate more information on other elements on land that are not cane, as they help the model distinguish between cane and other land covers. This exercise included some cassava and maize, correctly identified by the model, but with more examples, we will obtain better results. Specifically, we propose the inclusion of grasslands into the training data set. This is based on the fact that grass is from the same family as sugar cane, “Poaceae”, and the high probability of presence of grass areas in the region of study.
- Include among the model variables the altitude of the lot being cultivated, which determines its development.

More strategically, other useful actions for BAC can be:

- Apply this methodology to productive systems with similar phenologies and homologous growth habits (rice, cut pastures, maize, among others).
- Generate models for other productive systems. The data used in this project for maize and cassava is a good start.
- Retrain the model periodically, every six months, for example, since in any case, the visits are still carried out and information is collected in each of them

Finally, since cloud cover is one of the constant problems in the use of optical images, it is possible to consider the use of active sensor images, in particular Sentinel 1, which, being based on radar signals, do not present disadvantages with cloud coverage.

REFERENCES

- [1] Finagro. (Dec. 2020). *Manual de Servicios*. [Online]. Available: <https://www.finagro.com.co/manualserv/index.html>
- [2] *FINAGRO Modifica el Título Quinto de su Manual de Servicios*, NOTIBANCO Edición 2506, Banco Agrario de Colombia, Bogotá, Colombia, Jan. 2018.
- [3] Gerencia Nacional de Análisis Agropecuario, “Controles de inversión agropecuario,” Banco Agrario de Colombia, Bogotá, Colombia, Tech. Rep. CR-PR-006, Jan. 2020.
- [4] DANE. (2016). *Censo Nacional Agropecuario*. Accessed: Jun. 14, 2021. [Online]. Available: <https://www.dane.gov.co/files/CensoAgropecuario/entrega-definitiva/Boletin-12-UPNA/12-Boletin.pdf>
- [5] Ministerio de Agricultura y Desarrollo Rural. *Cadena Agroindustrial de la Panela. Bullets 2013*. (Dec. 2013). [Online]. Available: <https://sioc.minagricultura.gov.co/Panela/Documentos/2013-12-30%20Cifras%20Sectoriales.pdf>
- [6] (Oct. 2018). *Cadena Agroindustrial de la Panela. Bullets 2018*. [Online]. Available: <https://sioc.minagricultura.gov.co/Panela/Documentos/2018-10-30%20Cifras%20Sectoriales.pdf>
- [7] AgroNet. (Oct. 2014). *Caña Panelera*. [Online]. Available: <https://www.agronet.gov.co/Documents/Ca%C3%B1a%20Panelera.pdf>
- [8] T. Kurosu, M. Fujita, and K. Chiba, “The identification of Rice fields using multi-temporal ERS-1 C band SAR data,” *Int. J. Remote Sens.*, vol. 18, no. 14, pp. 2953–2965, Sep. 1997.
- [9] M. A. Peña and A. Brenning, “Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile,” *Remote Sens. Environ.*, vol. 171, pp. 234–244, Dec. 2015.
- [10] S. A. Kharat and V. B. Musande, “Cotton crop discrimination using Landsat-8 data,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 5, pp. 4381–4384, 2015.
- [11] J. Wang, X. Xiao, L. Liu, X. Wu, Y. Qin, J. L. Steiner, and J. Dong, “Mapping sugarcane plantation dynamics in Guangxi, China, by time series Sentinel-1, Sentinel-2 and Landsat images,” *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111951. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425720303217>
- [12] P. J. Murillo-Sandoval and J. A. Carbonell-González, “Principios y aplicaciones de la percepción remota en el cultivo de la caña de azúcar en Colombia,” Servicio de Cooperación Técnica y Transferencia de Tecnología, Cenicafé, Cali, Colombia, 2012.
- [13] E. Bastidas-Obando and J. Carbonell-Gonzalez, “Evaluating the applicability of MODIS data for forecasting sugarcane yields in Colombia,” in *Proc. Int. Soc. Sugar Cane Technol. (ISSCT)*, vol. 29, Durban, South Africa, 2007, pp. 1–4.
- [14] P. Murillo, C. A. Osorio, J. A. Carbonell, and A. E. Palma, “Monitoring sugarcane crops in the Cauca river valley (Colombia), using MODIS satellite images,” in *Proc. Int. Soc. Sugar Cane Technol.*, vol. 27, 2010, pp. 1–5.
- [15] P. J. Murillo-Sandoval, J. A. Carbonell-González, and C. A. Osorio-Murillo, “Evaluation of Landsat 7 ETM+ data for spectral discrimination and classification of sugarcane varieties in Colombia,” *J. Agricult. Sci. Technol.*, vol. 1, no. 5, pp. 101–107, 2011. [Online]. Available: <http://www.davidpublisher.com/Public/uploads/Contribute/55d43f18e9c5b.pdf>
- [16] A. Sakuma and H. Yamano, “Satellite constellation reveals crop growth patterns and improves mapping accuracy of cropping practices for subtropical small-scale fields in Japan,” *Remote Sens.*, vol. 12, no. 15, p. 2419, Jul. 2020.
- [17] D. Ashourloo, H. S. Shahrabi, M. Azadbakht, A. M. Rad, H. Aghighi, and S. Radiom, “A novel method for automatic potato mapping using time series of Sentinel-2 images,” *Comput. Electron. Agricult.*, vol. 175, Aug. 2020, Art. no. 105583.
- [18] P. Griffiths, C. Nendel, and P. Hostert, “Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping,” *Remote Sens. Environ.*, vol. 220, pp. 135–151, Jan. 2019.
- [19] E. Chuvieco, I. Aguado, M. Yebra, H. Nieto, J. Salas, M. P. Martín, L. Vilar, J. Martínez, S. Martín, P. Ibarra, J. de la Riva, J. Baeza, F. Rodríguez, J. R. Molina, M. A. Herrera, and R. Zamora, “Development of a framework for fire risk assessment using remote sensing and geographic information system technologies,” *Ecol. Model.*, vol. 221, no. 1, pp. 46–58, Jan. 2010.
- [20] E. S. Rodríguez, M. Á. T. Crespo, A. F.-P. Carmona, M. A. Alba, I. L. P. Serrato, and L. G. Ruiz, “Comparación del ndvi con el pvi y el savi como indicadores para la asignación de modelos de combustible para la estimación del riesgo de incendios en Andalucía,” in *Proc. Tecnologías Geográficas Para el Desarrollo Sostenible: Congreso del Grupo de Métodos Cuantitativo, SIG y Teledetección*. Alcalá de Henares, Spain: Universidad de Alcalá. Asociación de Geógrafos Españoles, Sep. 2000, pp. 164–174.

- [21] *Surface Reflectance Code (LaSRC) Product Guide*, USGS Landsat, USGS NASA, Reston, VA, USA, 2020.
- [22] USGS. (Aug. 2020). *Landsat 8 Collection 1 Land Surface Reflectance Code Product Guide | United States Geological Survey*. Accessed: Jan. 19, 2022. [Online]. Available: <https://www.usgs.gov/media/files/landsat-8-collection-1-land-surface-reflectance-code-product-guide>
- [23] ESA. (Mar. 2021). *Sentinel-2 Products Specification Document*. Accessed: Jan. 19, 2022. [Online]. Available: <https://sentinel.esa.int/documents/247904/685211/Sentinel-2-Products-Specification-Documents>
- [24] A. C. D. S. Luciano, M. C. A. Picoli, D. G. Duft, J. V. Rocha, M. R. L. V. Leal, and G. L. Maire, "Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm," *Comput. Electron. Agricult.*, vol. 184, May 2021, Art. no. 106063.



language processing, big data, recommender systems, and digital forensics.

GEOVANNY GONZÁLEZ-RODRÍGUEZ was born in Cúcuta, Norte de Santander, Colombia, in 1999. He received the degree in system and computing engineering from the Universidad de los Andes, Bogotá, Colombia, in 2021, where he is currently pursuing the degree in computer science. Since 2019, he has been a Teaching Assistant with the System and Computing Engineering Department, Universidad de los Andes. His research interests include computer vision, natural



Professor at the System and Computing Engineering Department, Universidad de los Andes. His research interests include network optimization from a multi-criteria perspective, smart mobility: vehicle networks and intelligent transport systems, the Internet of Things, and network security.

CARLOS LOZANO-GARZON (Senior Member, IEEE) received the degree in system engineering from the Universidad Nacional de Colombia, Bogota, Colombia, in 2002, and the Ph.D. degree from the Universidad de los Andes, Colombia, and the Universitat de Girona, Spain, in 2017. He worked as a Professor at the Universidad Nacional de Colombia, among others, and as a Visiting Professor at the Universidad Nacional de Asunción, Paraguay. He is currently an Assistant



DAVID NIÑO was born in Yopal, Casanare, Colombia, in 1996. He received the degree in system and computing engineering from the Universidad de los Andes, Bogotá, Colombia, in 2021, where he is currently pursuing the degree in computer science. He has been associated to the COMIT Research Group, Universidad de los Andes, since 2018, where he has been able to participate in remote sensing related projects. His research interests include computer vision, data analytics, and deep learning.



UASIG, the interest group on geographical information systems conformed by all the dependencies of the university concerned by GIS technology. His research interests include geographical information systems and knowledge management.

GERMÁN BRAVO-CÓRDOBA graduated in computing and system engineering from the Universidad de los Andes, Bogota, Colombia. He received the D.E.A. (M.Sc.) degree from the Institut National Polytechnique de Toulouse (INPT), Toulouse, France. Since 1989, he has been an Associate Professor at the Computing and Systems Department, Universidad de los Andes. He works in the information engineering track of the COMIT Research Group and coordinates



learning, data mining, natural language processing, knowledge engineering, and applied artificial intelligence, as well as published several papers in specialized journals and conferences.

HAYDEMAR NUÑEZ received the master's degree in computer science from the Central University of Venezuela (UCV), Caracas, Venezuela, the D.E.A. degree from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, and the Ph.D. degree in artificial intelligence from the UPC. She is currently a Visiting Professor at the Universidad de los Andes, Bogotá, Colombia. During her career as a Teacher and a Researcher has conducted projects in the areas of machine



ogy (COMIT) Research Group which main research focus are distributed systems and high performance computing (HPC) systems. He personally leads institutional and national cloud and HPC initiatives, and he was national coordinator for the establishment of a grid HPC platform between Europe and Latin America. His research interests include HPC, distributed systems, and cloud computing.

HAROLD CASTRO (Senior Member, IEEE) graduated in computing and system engineering from the Universidad de los Andes, Bogota, Colombia. He received the D.E.A. (M.Sc.) and Ph.D. degrees in computer science from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France. He is currently an Associate Professor at the Computing and Systems Department, Universidad de los Andes. He is also the Director of the Communications and Information Technol-



sensing, starting with the construction of the tagged data set and the implementation of machine learning algorithms for the classification of land cover. Her research interests include cloud computing, big data, machine learning, and recommender systems.

CAROLINA PARDO graduated in multimedia engineering from the Universidad Militar Nueva Granada, Bogotá, Colombia. She received the master's degree in information engineering from the Universidad de los Andes. She worked in the construction and coordination of proposals for projects focused on the area of geographic information systems (GIS) in education and humanitarian sectors. She has carried out projects focused on the analysis of satellite images through remote



AURELIO VIVAS graduated in computing and system engineering from the Universidad del Valle, Cali, Colombia. He received the M.Sc. degree from the Universidad de los Andes, Bogotá, Colombia, where he is currently pursuing the Ph.D. degree. Since 2018, he has been a Teaching Assistant with the System and Computing Engineering Department. He has been able to participate in remote sensing, desktop grid computing, and high-performance molecular dynamics projects. His research interests include programming languages, scientific parallel computing, and software-defined infrastructures.



LUIS CARLOS MOTTA received the degree in electronic engineering from the London College of Management and Technology and IT, and the master's degree from the University of East London. His research interests include IT, management, and digital transformation. He is the winner of Colombian's Inventor Award in Research Category in 2016.



YUBER CASTRO received the degree in system engineering from the Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, and the degree in system audit from the Universidad Católica de Colombia. He is currently pursuing the master's degree in data science and communications master with the Universidad Distrital Francisco José de Caldas and works at Banco Agrario de Colombia as University Professional in the area of Investment Monitoring and Control.



JULIO RENE ROJAS was born in Bogotá, Colombia. He received the degree in agricultural engineering from the Universidad Nacional de Colombia, Bogotá, in 2016. From 2014 to 2016, he was a Research Assistant at the Precision Agriculture Research Group, Universidad Nacional de Colombia. He has participated in national soil congresses as a speaker on zoning and management of saline soils. Since 2016, he has been working as a professional with the investment control area at Banco Agrario de Colombia, linked to the National Management of Agricultural Analysis, on issues of planning and information models aimed at developing technical agricultural strategies. His research interests include precision agriculture, sustainable development, and remote sensing.



JAZMIN MEDINA received the degree in systems engineering. She has work experience in consulting, formulation, structuring, and management of public investment projects in several sectors, such as agribusiness, STI, and health. She is a Specialist in process management and quality. She worked as an Advisor of science, technology, and innovation projects with the Universidad Surcolombiana. She currently works at Banco Agrario, supporting the proposals of STI projects from the digital innovation management.



LUIS IGNACIO SUÁREZ received the degree in system engineering, M.Sc. degree, and the degree in economics from the Universidad de los Andes, Bogotá, Colombia. In the last four years ago, he is professionally linked to Banco Agrario de Colombia. Since 2019, he assumed the challenge of becoming the Vice President of Digital Innovation with Banco Agrario de Colombia. In 2003, he was a Professor at the Pontificia Universidad Javeriana. In 2004, he was a Professor at the Universidad de los Andes. He has extensive experience in management positions, as well as advising and directing research projects related to information technologies.

...