COVERAGE CHANGES DETECTION AT CIÉNAGA GRANDE, SANTA MARTA – COLOMBIA USING AUTOMATIC CLASSIFICATION

J.S. Vinasco[1], D.A. Rodríguez[1], S. Velásquez [1], D.F. Quintero[1], L.R. Livni[1], F. L. Hernández[1,*]

[1] Remote Sensing Research Group, Universidad del Valle, Santiago de Cali, Colombia - (vinasco.juan, diego.rodriguez.mariaca, sergio.velazquez, duvan.quntero, lizeth.cuervo, francisco.hernandez)@correounivalle.edu.co

KEY WORDS: Automatic classification, Neural Network Analysis, Random Forest, Landsat 8, vegetation index, NDVI, EVI, NDWI

ABSTRACT:

The Ciénaga Grande, Santa Marta is the largest and most diverse ecosystem of its kind in Colombia. Its primary function is acting as a filter for the organic carbon cycle. Recently, this place has been suffering disruptions due to the anthropic activities taking place in its surroundings. The present study, the changes in the surface of Ciénaga Grande, Santa Marta, Magdalena, Colombia between 2013 and 2018 were determined using semiautomatic detection methods with high resolution data from remote sensors (Landsat 8). The zone of studies was classified in six kinds of surfaces: 1) artificial territories, 2) agricultural territories, 3) forests and semi-natural areas, 4) wet areas, 5) deep water surfaces & 6) wich is related to clouds as a masking method. Random Forest classifiers were utilized and the Feed For Ward multilayer perceptron neuronal network (ANN) was simultaneously assessed. The training stage for both methods was performed with 300 samples, distributed in equal quantities, over each coverage class. The semi-automatic classification was carried out with an annual frequency, but the monitoring was carried out throughout the analysis period through the performance of three indicators Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and Normalized Difference Water Index (NDWI). It was found from the confusion matrix that the Random Forest method more accurately classified four classes while Neural Networks Analysis (NNA) just three. Finally, taking the Random Forest results into account, it was found that the agricultural expansion increased from 7% to 9% and the urban zone increased from 20% to 30% of the total area. As well as a decrease of damp areas from 27% to 12% and forests from 4% to 3% of the total area of study.

1. INTRODUCTION

The most commonly used methods to study the changes of the vegetal coverages demand exhaustive field surveys which increase the costs of the projects and condition the spatial variable (Di Vittorio and Georgakakos, 2018), nevertheless, the evolution of the remote sensing has allowed the terrestrial ecosystems to be characterized (Setiawan et al., 2016) in a way that it is possible to quantify the changes on the surfaces, rate their state and therefore monitoring the spatial dynamics of the swamps (Wu et al., 2017).

One of the most conventional methodologies of remote sensing is the employ of standardized indexes. Faour et al., (2016) studied the change of the vegetation in the Middle East region by employing the Normalized Difference Vegetation Index (NDVI), classifying it into different categories which describe the behavior of the soil, concluding that arab soil suffers from humongous loss of vegetation. Othman et al. (2014) considered NDVI as one of the fittest methods to observe changes of the vegetation and NDWI for waters thanks to the results obtained after the study at Razzaza Lake, where it was possible to quantify and observe the increase and decrease of the vegetal coverage over time.

In addition, there are more sophisticated methodologies which allow to determine and distinguished the different kinds of coverage, such as classification with remote sensing data using supervised and unsupervised methods (Belgiu & Drăgu, 2016). Some supervised methods are based upon neural network learning (ANN), Classification and Regression Trees (CART). In addition, their assembled methods such as Random Forest have been useful for studying the distribution of the different coverages of a swamp, the changes it suffers during a determinate term and the quantification of these areas (Han et al., 2012).

Automatic classification methods have demonstrated high accuracy over satellite imaging (Wright y Gallant, 2007), as well as shorter processing terms and ability to efficiently estimate changes in the coverage compared to conventional techniques (Han et al., 2012); (Munyati, 2000). The use of ANN is one of the classification methods, a model based on a large amount of training data which later can predict to what kind each pixel belongs, achieving accuracies up to 93.4% as in the case of Yang et al. (2018). Han et al., (2012) compared different variations of the ANN method, among which are the radial-based neural networks with those that obtained an accuracy of 48 to 86%, and also compared the ANNs to RF, in which reached an accuracy of 96%.

The objective of the present study is to determine the changes in the surface of The Great Swamp of Santa Marta, Magdalena, Colombia between 2013 and 2018 by using Semi-automatic classification methods with high resolution from remote sensors. Conventional indexes of vegetation and water were estimated to determine their temporary courses and behaviors. Subsequently, two methods of supervised classification were employed to discriminate the types of water, vegetal and artificial coverage. The authors assessed the reliability of the models by using confusion matrixes generated by means the internal validation of each surface type. Finally, the changes were estimated by means of a simple differences model.

2. STUDY AREA AND MATERIALS

The study area is located in the Caribbean Region in the Department of Magdalena with coordinates al center N 10° 51' 51". The Great Swamp has an approximate extension of 5.000 km2, it has been classified as a zone of arid tropical weather with temperatures between 27°C and 30°C influenced by the Intertropical Convergence Zone, with annual average

precipitation values up to 807 mm (Freites et al., 2001) with a dry season between December and May and a rainy season from June to November, of which October presents the highest values. This area is categorized as rural land and it shelters the 64.7% and 50.5% of the population of the municipalities Pueblo Viejo and Sitio Nuevo, respectively; which develop fishing and tourism as their main economic activity, nevertheless, nearby zones to the municipality of Pueblo Viejo have implemented agriculture based on banana and African palm tree crops (Aguilera, 2011).

A set of 119 Landsat-8 images was utilized to cover the study term 2013 – 2018. The Landsat-8 Satellite offerts a total of eleven (11) bands (Table 1) generated by the two sensors on board: The Operational Land Imager (OLI) for the visible spectrum sensor and, the Thermal Infrared Sensor (TIRS) for the infrared thermal region. Despite existing other sources of spectral information with greater advantages in processing times (Setiawan et al., 2016); the 12-bit radiometric resolution, the temporary resolution that provides images every 16 days and finally, the spatial resolution of 30m x 30m makes it more convenient to use Landsat images for studies to estimate coverage changes (Haque & Basak, 2017).
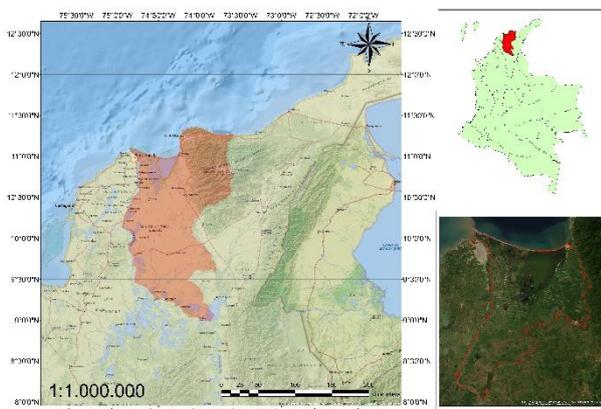
Fig. 1. Study area "Cienaga Grande, Santa Marta"

3. METHODOLOGY

3.1 Pre-Processing of Landsat-8 Images

Landsat-8 data was converted to planetary reflectance through the use of rescaled reflectance coefficients, arranged in the MLT metadata file, considering the observation angle in the digital value equation (ND) to reflectance (IGAC-CIAF, 2013).

$$\rho_{TOA} = \frac{M_\rho * Q_{CAL} + A_\rho}{sin(\phi se)} \quad (1)$$

where: ρ_{TOA} = reflectance value at the top of the atmosphere
Mρ = multiplicative factor for each band
Aρ = additive value for each band
Q_{CAL} = digital number of each pixel
$sin(\phi se)$ = solar elevation angle

Atmospheric correction DOS that subtracts all the minimum reflectance values from the minimum band values (Chavez, 1988) was applied to all 119 images converted into planetary reflectance and angularly corrected.

Two indicators were selected to highlight land coverage (NDVI, EVI) and water bodies' behavior (NDWI). The Normalized Difference Vegetation Index (NDVI) makes a classification of the terrestrial coverages to estimate the quantity, quality and vegetation health through the interpretation of photosynthetic processes (Rouse, et al., 1974). A healthy plant disperses solar radiation in the near-infrared region so as not to generate an overheating that would damage tissues and absorb solar radiation from the red region since it is part of an active photosynthetic radiation (Othman et al., 2014).

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \quad (2)$$

where: ρ_{NIR} = near wavelength infrared reflectivity (Landsat-8 band 5)
ρ_{RED} = red wavelength reflectivity (Landsat-8 band 4)

The Enhance Vegetation Index (EVI) provides information that allows monitoring the state of the vegetation in case of high biomass densities. The EVI was optimized from the NDVI to reduce the influence of the atmosphere called "aerosol resistance" used by the blue band to correct the effect of aerosols on the red band (Huete et al., 2002). The EVI was evaluated for characterizing NDVI's saturation situations, caused by the influence of intertropical convergence on the study area for some periods of the year.

$$EVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + C1 * \rho_{BLUE} - C2 * \rho_{BLUE} + L} \quad (3)$$

where: ρ_{BLUE} = blue wavelength reflectivity (band 2 Landsat 8)
L = ground adjustment factor (1.0)
$C1, C2$ = 4 band correction constant for atmospheric aerosol dispersion (6.0 and 7.5 respectively)

The Normalized Difference Water Index (NDWI) allows wet areas to be identified from maximizing the reflectivity of the water with the wavelengths of the green band and with the reflectivity generated by the near-infrared spectrum to minimize it., hence this relation takes leverage of the high reflectance of que Near Infrared (NIR) for terrestrial vegetation and soil characteristics (McFeeters, 1996).

$$NDWI = \frac{\rho_{GREEN} - \rho_{NIR}}{\rho_{GREEN} + \rho_{NIR}} \quad (4)$$

where: ρ_{GREEN} = green wavelength reflectivity (Landsat-8 band 2)

3.2 Supervised Classification

3.2.1 Random Forest

This technique associate to statistical and machine learning disciplines, has its matrix model as decision trees for classification and regression (CART). These are models that, from the statistics values, such as the mean, median, standard deviation, and Gini coefficient, find optimal values to classify the information. In cases of classification, the class for each category can be inferred based on a divided complex model. However, the CART method has multiple inconveniences that excluded it to be applied to real-life exercises, Among these, are their predisposition to bias and lack of robustness, because they are highly dependent on the distribution of the training set, which

implies that training two trees with the same initial conditions does not guarantee that the results are the same or even similar.

Therefore, different strategies have been proposed to attack this problem. Among them, two stand out: the bagging uses a set of random subsamples for the training of the different weak classifiers, attacking the bias of the individual CARTs, that is to say the different Trees specialize in various classes; and the impulse (boosting) proposes that a strong classifier can be generated by a weak one. Finally, the RF algorithm takes a set of training data and automatically divides it into subsamples. Then, 500 trees were independently train for the purpose of the study and to respond to a class, taking the most voted value from the 500 trees.

3.2.2 Artificial Neural Networks

They are an artificial intelligence model based on the human brain behavior, it is mainly based in neurons and their connectors. The models are a group of elements (neurons) that work together. Each part of the network receives information to be sent through the connectors to other neurons that allow them to know and learn from the data. There are many neural network models used for different activities, in the case of classifications, the perceptron is the simplest and most used neural network structure (Caicedo & López, 2009).

For this study, a multilayer perceptron neural network FeedForWard (MLP) type was used. This model used two hidden layers. One of these layers, has 7 neurons and a hyperbolic tangent activation function (Tanh). The other layer has 10 neurons and a linear exponential unit function (Selu). The activation function for the output layer, with 7 neurons, was the normalized exponential (Softmax), based on the logistic function. The Softmax function selection is due to its categorical distribution, commonly used for classifications (Caicedo & López, 2009).

$$f(z) = \frac{1}{1 + be^{-az}}$$
(5)

where: a = the exponential parameter
b = the exponential prefactor

Layer	Associated function	Number of neurons
Input		7
Hide 1	Hyperbolic Tangent	7
Hide 2	Exponential linear scaling	10
Output	Softmax	7

Table 1. Number of neurons and activation function for each layer

3.2.3 RF and ANN Sampling and Training

For the training of the machine learning technique, 300 samples were used distributed into six categories; five soil coverage taken from the adaptation of European methodology for classification CORINE Land Cover to Colombia in its most basic level and an additional category which corresponds to the presence of clouds in the scenario. The kinds obtained by CORINE refer to artificial, agricultural territories, forests and semi-natural areas, damp areas and water surfaces. The 300 samples fed the models of ANN and Random Forest; in both cases, 15% of the data was utilized to prove the validity of the adjusted models and the remaining 85%

fed the model. The root mean square and the accuracy associated with ANN were 0.25 and 0.95 respectively.

The sample size is the main affectation in the training of the ANN model. For this reason, Bolgiu & Drăgut (2016), recommend a sample of 0.25% of the size of the area, which corresponds to a sample size, for the case of the present study, of 8856 data.

4. RESULTS

The results of this work present a practical approach to the use of semi-automatic techniques in Colombia for the derivation of land cover maps, in contrast to the methodologies officially accepted by the national environmental authorities. Among them we have the one established by the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM, by his initials in Spanish) that adapted the European methodology of Corine Land Cover for National studies at 1: 100 000 scales (Ardila et al, 2010), in this adaptation, the determination of land cover is given through photointerpretation techniques which may include image enhancement techniques such as false color compositions, spatial filters, and fusion of multispectral bands with the panchromatic band on the Landsat satellite (IDEAM, I. (2008).

4.1 Vegetation and Water Indices

In Fig. 2, it is observed how the EVI values obtained for the crop and forest sample areas are within the range mentioned by Roldán & Poveda, (2006). In addition, it is evident that the behavior of the index, described Roldán & Poveda, (2006) for the Caribbean region, presents a similar pattern in its local minima, located between the months of November and April. On the other hand, the values obtained from NDVI present local minimums from January to April. In both cases, the indices present atypical values for some periods.

The trend pattern of the trajectories, as in the previous case, is consistent with the above for the Caribbean Region. In this case, the observed values are below those expected according to local characteristics (Roldán & Poveda, 2006). However, the values are within the average range of values described by other authors in previous studies (Huete et al., 2002).

In Ciénaga Grande de Santa Marta, two crops with similar crop characteristics are predominant, such as African palm and Banana (Aguilera, 2002). In the activity of harvesting the fruit there are no modifications in the appearance of the canopy, unless in the crops there is a presence of pests or extreme climatic conditions. Of which, for agricultural cover, the EVI and NDVI trajectories are represented with green and blue dashed lines (Fig. 2), however, they do not exhibit the representative phenological fluctuations of the surface. In addition, the NDVI index turns out to be very sensitive, so the depressions that apparently shows the cycles do not refer to crop cuts, but to the effects caused by precipitation.

The NDVI index presents a disturbance in the spectral response, so those depressions that apparently demarcate the cycles do not refer to crop cuts, but to what is considered by Faour et al. (2016), where they expresses that the index is affected by climatic behaviors, typical in the region. Therefore, it is observed that months of January to April spectral responses tend to take low values (stressed vegetation) and from June to November high values (vigorous vegetation).

The EVI index has a very stable pattern with an almost horizontal trend and without any abrupt change in the period analyzed, but this pattern is expected in a forested area that has not had any interruption due to anthropomorphic problems or a natural change. On the other hand, the NDWI index represented in dotted lines (Fig. 2) shows how deep-water bodies have a higher value compared to wetlands due to the proportion of the near and infrared green bands. NIR values are higher in wetlands due to

the turbidity that follows the increase in the equation. 4, shows that the difference of the numerator is smaller.
In addition, the trajectories are not able to reflect the phenology of the crops due to the scale of analysis. For this to happen, agricultural soils must have the same plantation and all crops should be synchronous with their sowing and harvesting date.
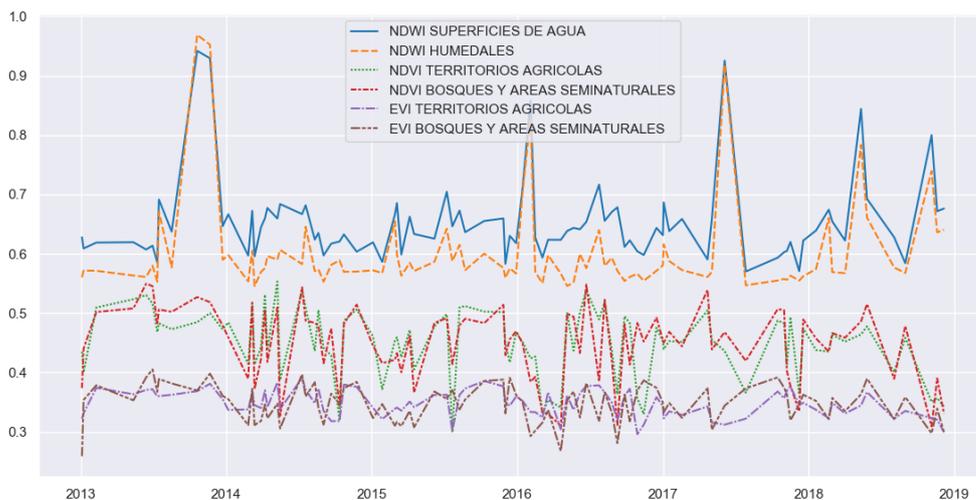
Fig. 2. Trend of the average values of the vegetation and water indices used

4.2 Classification Methods

Fig. 3 shows the RF confusion matrix, a 100% effectiveness is highlighted in four of the six categories (Urban Zone, Agricultural Territories, Water and Cloud surfaces), and an accuracy of 90% in the class assignment of wet areas. Forests and semi-natural areas were the classes with the lowest allocation accuracy (Fig. 3, left). Regarding the accuracy of the ANN,
it is observed that in three of the six categories, the allocation was 100% correct, for two categories the correct allocation was

higher than 80%, and in the classification of water surfaces, the correct allocation was 71% (Fig. 3, middle), which constitutes the lowest percentage of the assignments made through RF and ANN. In both methods, the assignment of urban area and clouds was done exactly. The water surface class presented greater discrepancy between the classifiers. Finally, the correlation analysis of the assignments performed by both methods show that most of the pixels were classified in the same class since the Pearson coefficient, statistically significant with p-value <0.01, was 0.81 (Fig. 3, right).
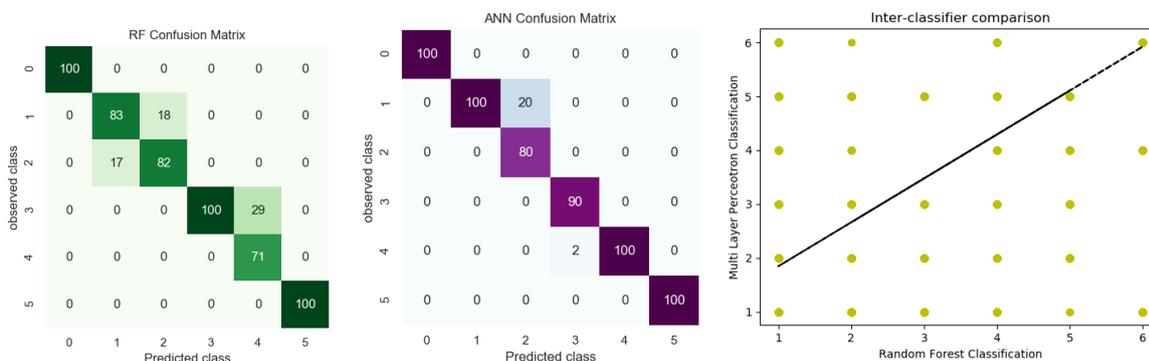
Fig. 3. Confusion Matrix and ANN and RF correlation scatter plot. 1) artificial territories. 2) agricultural territories. 3) forests and semi-natural areas. 4) wet areas. 5) water surfaces and 6) clouds

4.3 Coverage Changes Determination

After performing the overall fit and corresponding tests for both classification methods, the coverage change occurred during the

five years was determined using a difference model which was fed with the classification of September 2013 to 2018. It can be observed for the case of RF, a change of class from wet areas to water surface, as well as a decrease in forest areas (Fig. 4). In the ANN model, it is evident how an area of the swamp made a

transition from water surface to wet area. The presence of clouds and the edge of the swamp are observed in the extreme values of

ANN. In both cases, the classification of clouds has a high correspondence.
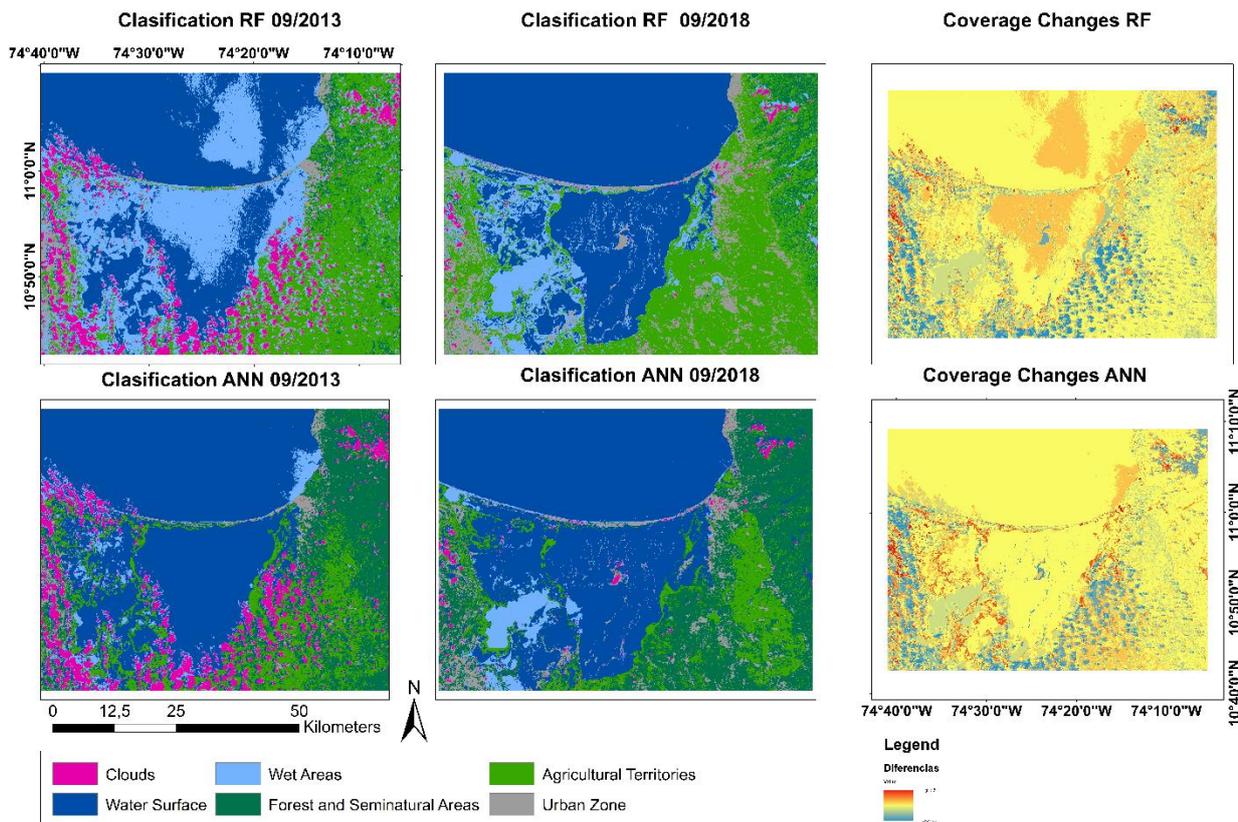
Fig. 4. Classification and changes in 2013 and 2018 coverage

Finally, an annual follow-up of the coverage change was carried out based on the classification results from both methods. It is observed how the RF method classifies more areas as agricultural territories instead of forests, and wet areas instead of water surfaces, compared to the ANN. Based on the statements made by Belgiu & Dragut., (2016) the RF method is better suited to the data in situations where they are fed with a small number of samples. In this sense, the results suggest that there are four important coverage changes: agricultural and urban expansion from 7 to 9%, and from 22 to 30%, of the total area, respectively. As well as a decrease in wet areas from 27 to 12%, and forests from 4 to 3% of the total area of the study area (Fig. 5).

Fig. 5. Percentage of coverage changes with respect to the total study area.

5. CONCLUSION

An evaluation of the ANN and RF classification methods was performed by comparing the success of each pixels. A correlation of 0.81 and a determination coefficient of 0.82 were obtained, which indicates that for 81% of the cases, both models match within the category assignment. However, the confusion matrix showed that the neural network model classified with high precision three of the six classes studied, while the RF model achieved four to 100%, which means that the RF model performed better in the classification process. Finally, it is worth noting that the study provided evidence about the coverage changes at the Ciénaga Grande, Santa Marta from 2013 to 2018.

In this study, the RF model shows better results than the neural network model. This can be mainly due to two reasons. The first is intrinsic to the models and corresponds to their ability to generate good results, their tendency to overtraining and generalization capacity. Although both models run the risk of over-adjustment due to the need to optimize the objective function using a greater number of parameters, so more knowledge is required to adjust the architecture of the models. On the other hand, the second reason is reduced to the training stage of the techniques, where the ANN model presents limitations in its results when the samples are limited, as in the present study.

The sample size established in the literature review indicated that it is necessary a 0.25% of the total area, which for this study

translates into 8865 samples. The sample sizes were the biggest limitations for the development of the study.

The time scale used is quite wide to be able to accurately predict small and detailed changes in crops that are related to the NDVI and EVI index, on the other hand, on top of water and wetland surfaces, outliers were observed that displaced the average generating peaks because samples were taken at sea and in flood areas that in dry seasons with moisture deficits.

It can not be possible to determine with exactitude the variability of the crops since the harvest technique does not generate changes in the coverings that allow to identify the state of the crop since the replacement of the plants is taking place gradually, as well as the harvest of the fruit.

6. REFERENCES

Aguilera, M., 2011. The economy of the swamps of the Colombian Caribbean. Bank of the Republic of Colombia.

Aguilera-Díaz, M. M. (2002). Palma Africana en la Costa Caribe : un semillero de empresas solidarias. Documentos de Trabajo Sobre Economía Regional y Urbana ; No. 30, 30. http://repositorio.banrep.gov.co/handle/20.500.12134/3172

Ardila, N. J. M., & García, U. G. M. (Eds.). (2010). Leyenda nacional de coberturas de la tierra: metodología CORINE Land Cover adaptada para Colombia: Escala 1: 100.000. IDEAM.

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Caicedo, F., López, J., 2009. A practical approach to artificial neural networks, Editorial Universidad del Valle. Santiago de Cali.

Chavez, P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sens. Environ. 24, 459–479. https://doi.org/10.1016/0034-4257(88)90019-3.

Di Vittorio, C.A., Georgakakos, A.P., 2018. Land cover classification and wetland inundation mapping using MODIS. Remote Sens. Environ. 204, 1–17. https://doi.org/10.1016/j.rse.2017.11.001

Freites, L., Uribe, E., Daniel, B., Sicard, L.Y.M.T., Investigaciones, C. De Mar, S.C., Col, B., Palo, P., Rita, D.S., Paz, L., 2001. Chapter 4, in: Wind and Wind Energy Atlas of Colombia. pp. 77–88.

Han, M., Zhu, X., Yao, W., 2012. Remote sensing image classification based on neural network ensemble algorithm. Neurocomputing 78, 133–138. https://doi.org/10.1016/j.neucom.2011.04.044.

Haque, M.I., Basak, R., 2017. Land cover change detection using GIS and remote sensing techniques: A spatio-temporal study on Tanguar Haor, Sunamganj, Bangladesh. Egypt. J. Remote Sens. Sp. Sci. 20, 251–263. https://doi.org/10.1016/j.ejrs.2016.12.003.

Huete, A., Didan, K., Miura, T., Rodriguez E., P., Gao, X., Ferriera L., G., 2002. Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Index. Remote Sens. Environ. 83, 195.

IDEAM, I. (2008). Mapa de Cobertura de la Tierra Cuenca Magdalena–Cauca: Metodología CORINE Land Cover adaptada para Colombia a escala 1: 100.000.

IGAC-CIAF, 2013. Descripción y Corrección de Productos Landsat 8 LDCM (Landsat Data Continuity Mission). Cent. Investig. y Desarro. en Inf. Geográfica del IGAC -CIAF. https://doi.org/10.5751/ES-06710-190329.

McFeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. https://doi.org/10.1080/01431169608948714.

Munyati, C., 2000. Wetland change detection on the Kafue Flats, Zambia, by classification of a multitemporal remote sensing image dataset. Int. J. Remote Sens. 21, 1787–1806. https://doi.org/10.1080/014311600209742.

Othman, A.A., Al-Saady, Y.I., Al-Khafaji, A.K., Gloaguen, R., 2014. Environmental change detection in the central part of Iraq using remote sensing data and GIS. Arab. J. Geosci. 7, 1017–1028. https://doi.org/10.1007/s12517-013-0870-0.

Roldán, P., Poveda, G., 2006. Spatio-temporal variabillty of the indices ndvi ano evi. with appllcaton to five colombian. Meterología Colomb. 10, 47–59.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. NASA special publication, 351, 309.

Setiawan, Y., Pawitan, H., Prasetyo, L.B., Lubis, M.I., Parlindungan, M., Nurdiana, A., 2016. Characterizing Spatial Distribution and Environments of Sumatran Peat Swamp Area Using 250 M Multi-temporal MODIS Data. Procedia Environ. Sci. 33, 117–127. https://doi.org/10.1016/j.proenv.2016.03.063.

Wright, C., Gallant, A., 2007. Improved wetland remote sensing in Yellowstone National Park using classification trees to combine TM imagery and ancillary environmental data. Remote Sens. Environ. 107, 582–605. https://doi.org/10.1016/j.rse.2006.10.019.

Wu, W. ting, Zhou, Y. xuan, Tian, B., 2017. Coastal wetlands facing climate change and anthropogenic activities: A remote sensing analysis and modelling application. Ocean Coast. Manag. 138, 1–10. https://doi.org/10.1016/j.ocecoaman.2017.01.005.

Yang, Z., Mu, X. D., Zhao, F. A. 2018. Scene classification of remote sensing image based on deep network grading transferring. Optik (Stuttg). 168, 127–133. https://doi.org/10.1016/j.ijleo.2018.04.092.