LAND USE PREDICTION ACCURACY OF DIFFERENT SUPERVISED CLASSIFIERS OVER AGRICULTURE AND LIVESTOCK ECONOMY-BASED MUNICIPALITY IN BRAZIL

João Lucas Della-Silva [1,2] , Tatiane Deoti Pelissari [3], *Daniel Henrique dos Santos [4], José Wagner Oliveira-Júnior [1,2], Larissa Pereira Ribeiro Teodoro [5], Amanda Jacobs Tavares [1] , Paulo Eduardo Teodoro [5], Dthenifer Cordeiro Santana [5] , Izabela Cristina de Oliveira [5], Fernando Saragosa Rossi [5] , Wendel Buêno Morinigo [1,2] and Carlos Antonio da Silva Junior [6]*

[1]Universidade do Estado de Mato Grosso (UNEMAT), Sinop, Mato Grosso, Brasil, amanda.jacobs@unemat.br, [2]Programa de Pós-Graduação em Biodiversidade e Biotecnologia da Amazônia Legal (PPG-BIONORTE), Sinop, Mato Grosso, Brasil, joao.lucas1@unemat.br, jrwagner.9803@gmail.com, [3] Universidade Estadual de São Paulo (UNESP), Programa de Pós-Graduação em Agronomia, Ilha Solteira, São Paulo, Brasil, deotipelissari@gmail.com, [4] Universidade do Estado de Mato Grosso (UNEMAT), Programa de Pós-Graduação em Biodiversidade e Agroecossistemas Amazônicos, Alta Floresta, Mato Grosso, Brasil, danielhsantos120@gmail.com, [5]Universidade Federal de Mato Grosso do Sul (UFMS), Departamento de Agronomia, Chapadão do Sul, Mato Grosso do Sul, Brasil, larissa.uems@gmail.com, eduteodoro@hotmail.com, dthennyfer.santana@hotmail.com, izabela.oliveira@unesp.br, fernandosarossi@gmail.com, [6]Universidade Estadual de Mato Grosso (UNEMAT), Departamento de Geografia, Sinop, Mato Grosso, Brasil, carlosjr@unemat.br.

ABSTRACT

The dynamics of land use and land cover (LULC) are of great importance for the management of natural resources, sustainable development and urban planning over geographic space, and this condition is sometimes supported by geoprocessing and remote sensing techniques. The decision for which classification typology presents the best results is related to the application and considering the LULC prediction as input to a cellular automata (CA) network, the performances of Classification and Regression Tree (CART), Random Forest (RF) and Minimum Distance (MID) for predicting land use and occupation in Sinop, Brazil were assessed. Using the median of the reference years 2013 and 2015 to create a transition potential modelling (TPM) neural network, and then predict a scenario in 2017. With the highest performance, the RF typology reached the best performance in an area of mostly agricultural occupation, separated into four classes (native forest, urban area, water and bare soil/agricultural activity).

Key words — Random Forest; Neural net; Decision tree; Landsat.

1. INTRODUCTION

Historical comparison of land use change (LULC) is very important for natural resource management and sustainable development, as it indicates the interaction between the environment and humans in that region [1]. For LULC classification, remote sensing techniques are often used as they significantly reduce the cost and time required for such task [2].

Remotely sensed data and the geoprocessing affords supervised learning methods, such as Artificial Neural Network (ANN) and Random Forest (RF), whom can be designed to solve classification problems [3].

Land use and land cover mapping are a useful tool in understanding and updating the use of land cover, playing a crucial role in planning and guiding decision-making [4]. In addition to the Amazon, we consider the importance of LULC dynamics over the Cerrado, which is a biome with a very rich biodiversity and is considered one of the world's hotspots for biodiversity conservation, which straddles about 23% of the Brazilian territory and borders the Amazon from south [5].

Tools such as the MOLUSCE plugin support the input of collateral data, which supports the classification stage in three periods, the first two for modeling the change detection algorithm, and a third for validation [6], closely related to the objective of this work. Therefore, to identify a LULC classifier with better performance for spatialized prediction from artificial neural networks, the objective of this study was to evaluate the accuracy of pixel-based classifiers Random Forest (RF), Classification and Regression Tree (CART) and Minimum Distance (MID) as input data for the MOLUSCE neural network prediction model.

2. MATERIAL AND METHODS

2.1. Study area

The study area relies on the municipality of Sinop, in the state of Mato Grosso, Brazil. It is in the north of the state, between latitudes 11°20′S and 12°10′S, and longitudes 55°50′W and 55°10′W, with an average altitude of 384 m (Figure 1). Classified as ‘Aw’ according to the Köppen-Geiger climate classification, the municipality has an average annual temperature of 27 °C [7].

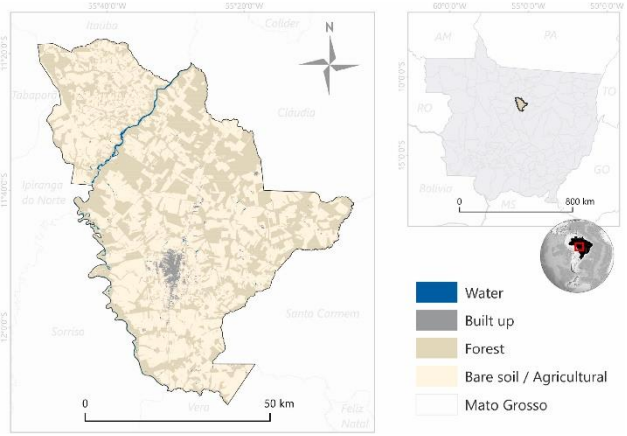


Figure 1. Location of Sinop, in the state of Mato Grosso, Brazil. The seconded area of the thematic map is based on Random Forest classification of 2017 median surface reflectance images.

2.2. Data acquisition

This study is based on median values from an annual image collection of the OLI/Landsat-8 time series over the study area from the years 2013, 2015 and 2017, acquired and calculated on the Google Earth Engine (GEE) platform already corrected for the top of the atmosphere. The image collection adopted is from the LANDSAT/LC08/C02/T1_TOA database. The definition of these years considered the use of the same orbital sensor and little change in the water bodies in the area. The Teles Pires river suffered flooding due to the insertion of a hydroelectric power plant in 2018, and this sudden change would limit the performance of the classifiers and the prediction model used.

The land use prediction through MOLUSCE reaches greater results with collateral data inclusion. For that, geomorphological inputs were based on digital elevation model of ALOS PALSAR data, based on the median value for 2017. The related data of aspect and slope were based in the very same raster data, processed in QGIS 3.24 software.

2.3. Land use classification

To verify the adherence of the land use and occupation classification models in relation to the specialized prediction of this same classification, a collection of images was created in the Google Earth Engine (GEE). Through GEE, the separation of reference samples into two data sets is carried out automatically, with the first set being used for classification (or training) and structuring the neural network (80% of the samples), while the second for model validation created in training and evaluate its accuracy (20% of samples).

2.4. Land use prediction

As a LULC prediction model, the MOLUSCE (Modules for Land Use Change Evaluation) on QGIS 2.18.24 plugin was applied (NextGIS and Asia Air NextGISAsia Air Survey, 2014). This plugin was developed to design a prediction model based on artificial neural networks (ANN), multilayer perceptron (MLP) typology, and for analysis and future simulation of LULC in simulation using ANN of the cellular automata (CA) type. In this plugin, LULC maps classified into two periods are inserted, with a view to predicting a third period equidistant in relation to the difference in input years.

The CA-based simulation model [8] seeks to relate the spatial arrangement of cells (i.e., pixels), their classification and establishment of a transition rule in the expansion of classes based on a previous scenario and with a time step defined from the interval between this initial state and the current state, the which presents a very simplified dynamics of the urbanization process in this type of application Von Neumann [8].

Using the plugin, several processes are developed with a view to quantitative and spatialized description of the change in LULC over the years. Distributed in stages, (I) the correlation between total areas per class, (II) the total area per class, (III) the difference in area per class between the years of entry are calculated, from which (IV) the model is constructed of potential change and (V) the simulation of the future scenario, as well as (VI) the validation of this predicted scenario against the actual classification.

3. RESULTS

Using the MOLUSCE plugin, total area values per class and their percentage value were calculated (**Table 1**). The prediction algorithm simulation involves class difference detection for each classifier, i.e., the class change is detected in each pixel for the 2013 and 2015 classifications.

Classifier	Class[1]	2013	2013%	2015	2015%	2017	2017%	Pred	Pred%
CART	1	2823.89	0.71	11198.02	2.81	3412.78	0.86	11439.81	2.87
	2	355.44	0.09	632.57	0.16	604.94	0.15	1138.15	0.28
	3	170560.18	42.73	133656.72	33.49	157311.68	39.41	133365.56	33.42
	4	225345.32	56.47	253597.52	63.54	237755.43	59.58	253141.31	63.43
MID	1	3162.84	0.79	2989.90	0.75	3009.37	0.75	2990.95	0.75
	2	35990.17	9.02	15243.37	3.82	20499.92	5.14	15241.97	3.82

	3	188719.17	47.29	172810.52	43.30	181272.18	45.42	172988.81	43.35
	4	171212.65	42.90	208041.04	52.13	194303.36	48.69	207863.10	52.08
RF	1	3124.08	0.78	3581.86	0.90	3118.29	0.78	3580.90	0.90
	2	5336.62	1.34	6129.32	1.54	6166.33	1.54	6127.92	1.53
	3	169344.24	42.43	152774.52	38.28	159866.07	40.06	152770.84	38.28
	4	221282.87	55.45	236602.11	59.29	229937.12	57.62	236608.16	59.29

[1] Water (1), Urban (2), Forest (3) and Agricultural / Bare soil (4).

Table 1. Measured (hectares) and relative (%) areas by class of classified years.

The design of a predicted LULC scenario rely on the transition potential modelling (TPM), a multilayer perceptron neural network (MLP) which includes verification of the global accuracy variation (ΔOA), minimum total error of validation (minimal validation overall error) and current kappa validation (current kappa validation).

Land use classifications on this study considered 2013, 2015 and 2017, with the last year being classified to validate each classification model in relation to the simulation from MOLUSCE (Figure 2). Right away, some classification errors can be seen for those results. CART classification over the whole time series present omission errors on ‘urban’ land use class. In contrast, several commission errors with the very same class are results for Minimum Distance. Our methods considered the potential classification errors with minor supervision, what may limit their performance in land cover prediction through cellular automata modeling.

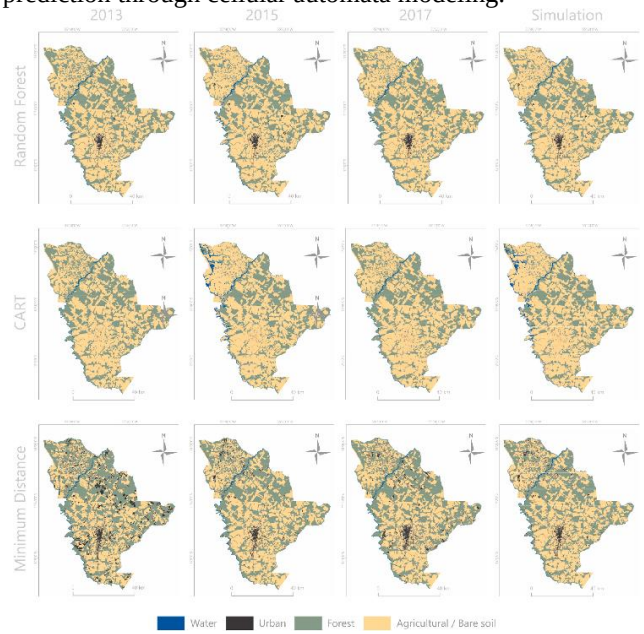


Figure 2. Classification of land use according to the Random Forest, CART and Minimum Distance classifiers for the municipality of Sinop, Mato Grosso.

In the validation stage with the classified scenario, the percentage of accuracy, the kappa statistics (total, location and histogram) were calculated, with different statistical approaches being sensitive to the location and in the form of a histogram of all classes, respectively [9]. Such results shows that all classifiers have a high level of accuracy, where the RF algorithm achieved the highest accuracy (0.96).

4. DISCUSSION

RF classification was efficient in differentiating the four classes, generating a more assertive map than the other classifiers. The CART classifier, in 2015, reduced performance for the 'water' class, and fail in classifying important part of 'Forest' class as 'water', and almost completely suppressed the class defined as 'Urban' over the four stages. Regarding the Minimum Distance results, the ‘Urban’ class was overestimated beyond the urban limits throughout the study area in the four periods, but specially for 2013 what affected the prediction and its classification accuracy.

Looking at Δ Overall Accuracy, RF classifier maintained its performance in the prediction model, (−0.0203) considered moderate in the literature [10]. This is due to the prediction model, which seeks to reproduce the maps of the respective classifieds as close as possible to the results generated by each one [11].

For the RF, a smaller variation was observed between the classifications of the reference years with other classifiers, which consequently infers greater accuracy in relation to the prediction.

Based on the cellular automata theory, the model for predicting the dynamics of land use and cover (LULC) was built, defined as a potential transition model (TPM) for the LULC data from the three classifiers considered in this evaluation. From this, the error and the number of training iterations of the multilayer perceptron network established for such spatialized prediction were defined, i.e., the learning curve (Figure 2), from which different patterns of the three learning curves were observed.

Finally, the definition of four classes of land use for a municipality with an agricultural economic base, such as Sinop, based on surface reflectance data from the OLI/Landsat-8 system presented satisfactory classification and validation results, as well as for the next prediction

process through MOLUSCE. Among the classifiers, we highlight the good performance of the Random Forest typology.

5. CONCLUSIONS

The performance comparison of classifiers for predicting the dynamics of land use allows the adoption of a model with better performance, which potentially permeates new analyzes relating land occupation to new scientific perspectives. The ground truth closely related to the classification process itself suggests Random Forest as the best classifier for this. Random Forest showed better performance, notably seen through accuracy and Kappa statistics. Our results grounds the use of Random Forest classifier as input for land use prediction through cellular automata artificial neural network, especially in the very same conditions of land uses classes and imagery.

8. REFERENCES

[1] O. I. Al Balasmeh and T. Karmaker. Effect of Temperature and Precipitation on the Vegetation Dynamics of High and Moderate Altitude Natural Forests in India. *journal of the indian society of remote sensing*, v. 48: pp. 121–144, Jan. 2020.

[2] A. Ghosh, R. Sharma, and P. K. Joshi. Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion. *applied geography*, v. 48: pp. 31–41, Mar. 2014.

[3] M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *isprs journal of photogrammetry and remote sensing*, v. 114: pp. 24–31, Apr. 2016.

[4] Z. Deng, X. Zhu, Q. He, and L. Tang. Land use/land cover classification using time series Landsat 8 images in a heavily urbanized area. *advances in space research*, v. 63: pp. 2144–2154, Apr. 2019.

[5] R. D. Françoso, R. Brandão, C. C. Nogueira, Y. B. Salmona, R. B. Machado, and G. R. Colli. Habitat loss and the effectiveness of protected areas in the Cerrado Biodiversity Hotspot. *natureza e conservacao*, v. 13: pp. 35–40, 2015.

[6] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *ieee journal of selected topics in applied earth observations and remote sensing*, v. 13: pp. 6308–6325, 2020.

[7] C. A. Alvares, J. L. Stape, P. C. Sentelhas, J. L. De Moraes Gonçalves, and G. Sparovek. Köppen's climate classification map for Brazil. *meteorologische zeitschrift*, v. 22: pp. 711–728, 2013.

[8] J. Von Neumann. The general and logical theory of automata, Cerebral Mechanisms in Behavior. The Hixon Symposium. *new york: john wiley & sons*, 1951.

[9] D. Gómez and J. Montero. Determining the accuracy in image supervised classification problems. *proceedings of the 7th conference of the european society for fuzzy logic and technology, eusflat 2011 and french days on fuzzy logic and applications, lfa 2011*, v. 1: pp. 342–349, 2011.

[10] J. Sim and C. C. Wright. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *physical therapy*, v. 85: pp. 257–268, 2005.

[11] C. Mozumder, N. K. Tripathi, and C. Losiri. Comparing three transition potential models: A case study of built-up transitions in North-East India. *computers, environment and urban systems*, v. 59: pp. 38–49, Sep. 2016.