

# LAND COVER CLASSIFICATION OF AN AREA SUSCEPTIBLE TO LANDSLIDES USING RANDOM FOREST AND NDVI TIME SERIES DATA

*Tatiana Dias Tardelli Uehara, Anderson Reis Soares, Renata Pacheco Quevedo, Thales Sehn Körting, Leila Maria Garcia Fonseca, Marcos Adami*

General Coordination of Earth Observation - OBT  
Brazil's National Institute for Space Research (INPE)  
Av. dos Astronautas, 1758, São José dos Campos - SP, Brazil  
(tatiana.uehara, anderson.soares, renata.quevedo, thales.korting, leila.fonseca, marcos.adami)@inpe.br

## ABSTRACT

Landslides are a natural, gravity driven phenomena which can cause great economic and human losses. To prevent them, Land Use and Land Cover (LULC) maps are essential to identify areas of high susceptibility and to detect landslide scars. This paper presents results of a classification of a landslide susceptible area, using Random Forest algorithm and time series. The time series dataset is composed by the Normalized Difference Vegetation Index (NDVI) values and 16 metrics derived from the time series. The best performance was achieved using 14 metrics plus the NDVI values, with overall accuracy of 93.23% and kappa equals to 0.8937. The metrics revealed a great capability for landslides detection.

**Index Terms**— landslide, time series, Random Forest, land cover, disasters

## 1. INTRODUCTION

Landslides are a type of mass movement, which is a widespread natural phenomenon that can be triggered by extreme precipitation events, earthquakes, snow melting or anthropic activities. This phenomenon can be considered a hazardous event when it occurs in a populated area. In Brazil, the landslides events were responsible for the second highest number of deaths between 1991 and 2012 [1].

In order to reduce the landslide susceptibility, inventory and LULC maps are crucial information. The inventory can help to understand the pattern of the events, while the LULC maps can assist with the areas that have experienced anthropogenic interventions that may induce landslides. Land use

---

We thank the National Council for Scientific and Technological Development (CNPq) for the financial support of this research, and the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES), and the Foundation for Science, Technology and Space Applications (FUNCATE) no. 17.2.0536.1 (Brazil Data Cube project), and grant 2017/24086-2, São Paulo Research Foundation (FAPESP). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

maps present a dynamic characteristic having the necessity to be constantly updated. Time series have been widely used to LULC mapping and to identify the nature of land cover changes [2]. Medium spatial resolution time series data have demonstrated high capacity for characterizing environmental phenomena, describing trends as well as discrete change events [3].

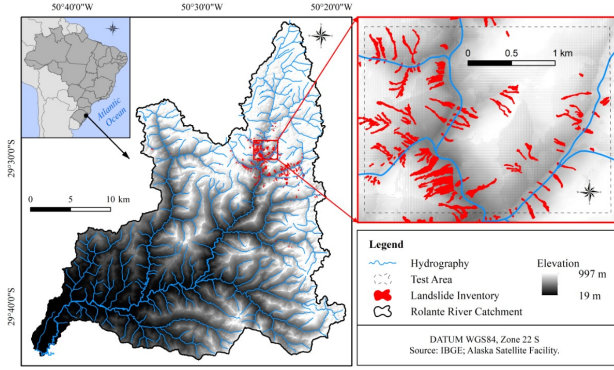
In this context, the aim of this research was to realize the LULC classification for an area with high concentration of landslides at the Rolante River Catchment using a Random Forest (RF) classification algorithm. We evaluated the performance of the NDVI and time series metrics for the classification and the landslides detection.

## 2. STUDY AREA

The study area has approximately 6 km<sup>2</sup> and is located at the Rolante River Catchment in the South of Brazil (Figure 1). The studied basin is placed between the latitudes of 29°20' - 29°43' S and the longitudes 50°18' - 50°43' W and its drainage area is 828 km<sup>2</sup>. The altitudes vary from 19 to 997m. The lithology is predominantly composed of basaltic rocks and sandstone and the climate is very humid subtropical. Numerous landslides occurred in this region on January 5th, 2017, caused by a torrential rain event [4].

## 3. METHODOLOGY

Sentinel 2A imagery was chosen, especially because it provides free orthorectified reflectance products with 10m of spatial resolution and a 5-day temporal resolution. Level 2A (Bottom-of-Atmosphere) product was not available for the day of the event, so Level 1C was used. The European Space Agency (ESA) provides an orthoimage product in Top-Of-Atmosphere (TOA) reflectance. In order to realize the atmospheric correction, the AtmosPy system [5] was used. It is a semi-automatic system of atmospheric correction which applies the 6SV (Second Simulation of a Satellite Signal in the



**Fig. 1.** Study area at Rolante River Catchment (RS/Brazil).

Solar Spectrum Vector version) model.

The time interval covered was from November of 2015 to September of 2019, with approximately one image per month. Images with cloud coverage above 60% have been removed. In total, the dataset was composed of 26 dates, which have been smoothed with an outlier removal filter. From each date of the Sentinel's image, the NDVI was extracted. This index presents a drastic reduction in its value when the landslide occurs. Besides that, once it is a composition of bands, it allows the use of the red and near-infrared bands in one single feature. This reduction of attributes is specifically beneficial when dealing with dense time series. The literature shows many successful researches using the NDVI to land cover mapping[6, 7].

The *tmetrics*<sup>1</sup>, package available in python<sup>1</sup>, was used to extract features from the NDVI time series. The package currently possess two modules to feature extraction. The first module computes basic metrics, that are derived from the time series using common statistical approaches. Currently, 7 metrics are available: mean, maximum, minimum, standard deviation, sum, amplitude and first slope.

The second module implements the metrics proposed by [8]. Currently, 9 metrics are available: area, area of the quadrant (AQ 1, AQ 2, AQ 3, AQ 4), circle, gyration, polar balance and angle. These metrics are derived from a time wheel legend proposed by [9]. To compute these features each time series has its values projected to angles in the interval  $[0, 2\pi]$ . A time series is a function  $f(x, y, T)$  where  $(x, y)$  is the spatial position of a point, and  $T$  is a time interval  $t_1; \dots; t_N$ , and  $N$  is the number of observations. The time series can be visualized by a set of values  $v_i \in V$ , where  $v_i$  is a possible value of  $f(x, y)$  in time  $t_i$ . Therefore, its polar representation is defined by a function  $g(V)\{A, O\}$  ( $A$  corresponds to the abscissa axis in the Cartesian coordinates, and  $O$  to the ordinate axis) where:

$$a_i = v_i \cos\left(\frac{2\pi i}{N}\right) \in A, i = 1, \dots, N. \quad (1)$$

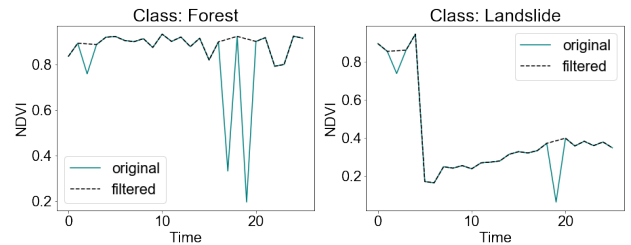
<sup>1</sup>github.com/andersonreisoares/tmetrics

and

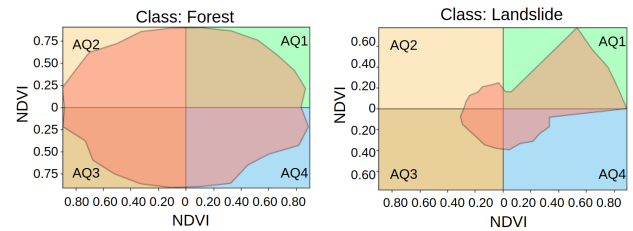
$$o_i = v_i \sin\left(\frac{2\pi i}{N}\right) \in O, i = 1, \dots, N. \quad (2)$$

In both equations,  $\frac{2\pi i}{N}$  is an arbitrary angle that depends on the acquisition date and  $v_i$  is the corresponding time series value. Considering that  $a_{n+1} = a_1$  and  $o_{n+1} = o_1$ , a closed shape is obtained.

Figures 2 and 3 illustrate an example of a linear and a polar projection of the NDVI for classes Forest and Landslide along the time series. For most of the classes considered in the classification, their main characteristic is a stable response along the period, except for Landslide, where an abrupt change occurs in the NDVI values.



**Fig. 2.** Linear projection of NDVI values in the time series for the classes Forest and Landslide. The blue line is the raw data, while the black dashed line, the filtered data. The time axis corresponds to the 26 dates used.



**Fig. 3.** Polar projection of NDVI values in the filtered time series for the classes Forest and Landslide. Starting from the green box lower right position, the time wheel goes anticlockwise.

A stratified proportional sampling was realized, being the proportion for each class based on previous classification works of the area [10]. As suggested by [11], high resolution imagery from Google Earth, was used as ground-truth for sampling. The samples are pixel based and were divided respecting the proportion 20/80; 133 testing and 532 training samples. The classes are: Agriculture, Landslide, Forest, Pasture and Bare soil.

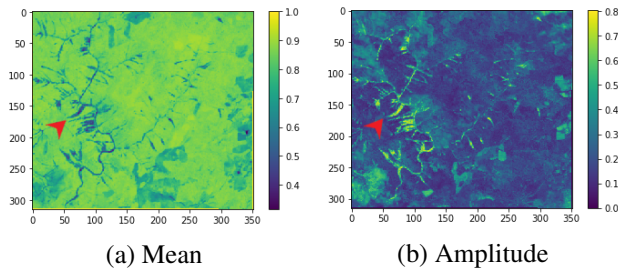
The classification was performed by a RF algorithm in python environment. The RF consists of a combination of decision tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class

to classify an input vector [12]. Many studies have proved the effectiveness of this classifier for LULC mapping [13, 14]. Four different groups of attributes were used in the classifications experiments: one single NDVI image, corresponding to the most recent date; the NDVI time series; the NDVI time series with the metrics and only the metrics.

#### 4. RESULTS AND DISCUSSION

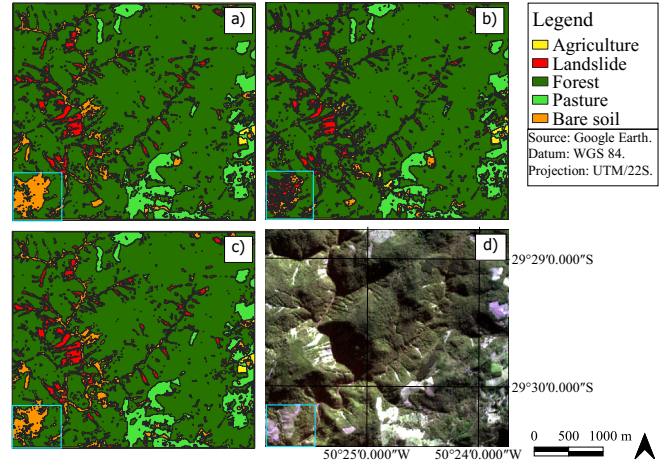
The first attempt to realize the LULC mapping via RF took into account a single NDVI image from September, 2019. The unsatisfactory results (overall accuracy of 66.92% and kappa index equals to 0.4731) showed that one single NDVI image was not enough to perform a good classification. Considering that, the NDVI time series composed of 26 images was used. The result showed an overall accuracy of 92.48% and a kappa index of 0.8827.

The metrics derived from the time series revealed a great potential to rapidly identify some important elements by visual interpretation. The features mean and amplitude assist the detection of the landslide scars, which is illustrated by Figure 4. Once the landslide scars correspond to not vegetated areas, hence presenting low NDVI values, they are identified in the mean attribute by their low values. Regarding the amplitude, which corresponds to the difference of the highest and the lowest values in the time series, the landslide scars present high values, once there is a significant change of the NDVI after the forest removal by the landslide. Polar metrics, also presented relevant material to the classification. As Figure 3 exemplifies, for Landslide class, AQ1 values are high, being remarkably reduced from AQ2 on. Whereas for Forest these metrics remain high and stable along the time series.



**Fig. 4.** Examples of basic features extracted. The red arrow points to one of the landslide scars.

The classification resulted from the use only of the metrics showed an satisfactory product with overall accuracy above 90%. Once the RF algorithm ranks the feature importance to the classification, the least important attributes were removed one at a time, then performing new classifications. The best result was achieved by removing the features minimum and maximum, keeping the following metrics: mean, standard deviation, sum, amplitude, first slope, area, AQ 1, AQ 2, AQ 3,



**Fig. 5.** Classification products. a) NDVI with metrics. b) Metrics. c) NDVI. d) RGB composition. The blue rectangle highlights the difference among the classification products.

AQ 4, circle, gyration, polar balance and angle. The product presented an overall accuracy of 92.48% and a kappa index equals to 0.8819. Comparing to the NDVI classification, the results for only the metrics showed more noise, producing small scattered polygons.

The last experiment united the NDVI with the above selected basic and polar metrics, using a total of 40 attributes. The result was superior from the other classifications, presenting an overall accuracy of 93.23% and a kappa index of 0.8937. Figure 5 illustrates the last three cited classifications. Special attention is given to the detached area in the blue rectangle. According to the ground truth reference, this area used to be forested, however an anthropic driven devastation transformed it, mainly, in bare soil and pasture. The product from the NDVI time series could correctly classify the area as bare soil. However, it's interesting to notice that the metrics product classified it, mainly, as landslide. That occurred, because the radical change in the NDVI value due to the devastation is similar to the landslide phenomenon, once it also presents an abrupt reduction in the vegetation index.

Table 1 illustrates the confusion matrix regarding the classification using the NDVI with the metrics. Bare soil class presented the highest comission and omission errors, being mistaken for Landslide and Forest. Once the sampling was proportional and stratified, this class had one of the lowest number of samples, which provoked the increased values of error percentage. One reason for mistaking this class with landslide is the extremely low NDVI values. In this case, a digital elevation model could help to distinguish both classes. Landslide class showed the second highest error values, with comission error equals to 25%. Forest presented low error values, however only Pasture and Agriculture showed 0% of comission and omission errors.

**Table 1.** Confusion matrix from the classification of the NDVI with the metrics.

Predicted \ True label	Agriculture	Landslide	Forest	Pasture	Bare soil	Comission error
Agriculture	<b>8</b>	0	0	0	0	0%
Landslide	0	<b>15</b>	3	0	2	25%
Forest	0	0	<b>71</b>	0	1	1%
Pasture	0	0	0	<b>23</b>	0	0%
Bare soil	0	2	1	0	<b>7</b>	30%
Omission error	0	11.76%	5%	0%	30%	<b>124</b>

## 5. CONCLUSIONS

The RF classifier presented a good performance for LULC mapping and identifying landslides. Except for the classification with only one single image as input, all other three experiments showed very satisfactory results, with overall accuracy above 90% and kappa index above 0.88. Even though the best classification result derives from the use of the NDVI time series with the metric features, no affirmation regarding this superiority can be done without the realization of statistical tests. These tests could prove if this improvement in the result is in fact significant or not. Furthermore, the metrics have shown a great potential to classification and, specially, to the visual detection of landslide scars.

Moreover, the NDVI as the unique feature for classification showed a great potential for characterizing landslides scars if more than one date is considered. This is due to the radical change in its values when the landslide event happens. For future studies, it is recommended to realize the above cited statistic tests, as well as trying experiments including a DEM, other sensor bands and vegetation indices.

## 6. REFERENCES

- [1] Centro Universitário de Estudos, “Atlas brasileiro de desastres naturais 1991 a 2012: volume brasil,” *CEPED. UFSC, Florianópolis*, 2013.
- [2] Claudia Paris, Lorenzo Bruzzone, and Diego Fernández-Prieto, “A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2019.
- [3] Cristina Gómez, Joanne C White, and Michael A Wulder, “Optical remotely sensed time series data for land cover classification: A review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55–72, 2016.
- [4] Renata Pacheco Quevedo, Laurindo Antonio Guasselli, Guilherme Garcia Oliveira, and Luis Fernando Chimelo Ruiz, “Modelagem de áreas suscetíveis a movimentos de massa: Avaliação comparativa de técnicas de amostragem, aprendizado de máquina e modelos digitais de elevação,” *Geosciences*, vol. 38, no. 3, pp. 781–795, 2019.
- [5] Felipe Menino Carlos, Vitor de Souza Martins, and Cláudio Clemente Faria Barbosa, “Sistema semi-automático de correção atmosférica para multi-sensores orbitais,” in *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*. Galoá, 2019.
- [6] D Jeevalakshmi, S Narayana Reddy, and B Manikiam, “Land cover classification based on ndvi using landsat8 time series: a case study tirupati region,” in *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2016, pp. 1332–1335.
- [7] Fanjie Kong, Xiaobing Li, Hong Wang, Dengfeng Xie, Xiang Li, and Yunxiao Bai, “Land cover classification based on fused data from gf-1 and modis ndvi time series,” *Remote Sensing*, vol. 8, no. 9, pp. 741, 2016.
- [8] Thales Sehn Körting, Leila Maria Garcia Fonseca, and Gilberto Câmara, “Geodma—geographic data mining analyst,” *Computers & Geosciences*, vol. 57, pp. 133 – 145, 2013.
- [9] R. Edsall, Alan MacEachren, and Donna Peuquet, “Assessing the effectiveness of temporal legends in environmental visualization,” 01 1997.
- [10] Tatiana Dias Tardelli Uehara, Sabrina Paes Leme P Correia, Renata Pacheco, Thales Sehn Körting Quevedo, Luciano Vieira Dutra, and Camilo Daleles Rennó, “Classification algorithms comparison for landslide scars,” *GEOINFO, 20 Years After!*, p. 158, 2019.
- [11] Pontus Olofsson, Giles M Foody, Martin Herold, Stephen V Stehman, Curtis E Woodcock, and Michael A Wulder, “Good practices for estimating area and assessing accuracy of land change,” *Remote Sensing of Environment*, vol. 148, pp. 42–57, 2014.
- [12] Leo Breiman, “Random forests,” *UC Berkeley TR567*, 1999.
- [13] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, and Gérard Dedieu, “Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas,” *Remote Sensing of Environment*, vol. 187, pp. 156–168, 2016.
- [14] Chuang Wang and Lizhen Lu, “Object-based random forest classification for detecting plastic-mulched landcover from gaofen-2 and landsat-8 oli fused data,” in *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. IEEE, 2019, pp. 1–5.