Random forest classification applied to sand pit detection in Cruzeiro do Sul, Acre, Brazilian Amazon

David Guimarães Monteiro França [1,2]
Liana Oighenstein Anderson [1,2]
Lucena Rocha [3]
Sacha Maruã Ortiz Siani[2]
Monique Rodrigues da Silva Andrade Maia [2,4]

[1] National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)
Estrada Dr. Altino Bondesan, 500, 12247-016, São José dos Campos – SP, Brazil
dvdgmf@gmail.com, liana.anderson@cemaden.gov.br

[2]Tropical Ecosystems and Environmental Sciences Group (TREES)
National Institute for Spatial Research (INPE)
Av. dos Astronautas, 1758, 12227-010, São José dos Campos – SP, Brazil

[3] Federal University of Acre - Forest Campus (UFAC)
BR 364, km 04, 69915-900, Rio Branco – AC, Brazil
lurubita@gmail.com

[4]Instituto Nacional de Pesquisas da Amazonia/CDAM
Instituto Nacional de Pesquisas da Amazônia, Coordenação de Dinâmica Ambiental.
Avenida Ephigenio Salles 2239 – Aleixo - 69000000 - Manaus, AM - Brasil
moniquerds@gmail.com

Abstract. Sand pit extraction, known as "Canchas", is recognized as one of the major environmental problems in northern Acre. Canchas impact the environment by not allowing the native vegetation to regenerate but also polluting water springs in the nearby areas. This paper has two objectives. First, we test the use of the Sentinel-2 data to detect sand pits, which usually are relatively small areas, with borders with water bodies and dense vegetation. Secondly, by using the random forest algorithm and multiresolution segmentation procedure combined with data mining, we test the generalization of this method by applying the settings defined in the first objective to an independent dataset (image) for the same location. The study area is located in the municipality of Cruzeiro do Sul, northwest portion of Acre state, Brazilian Amazonia. Preliminary results show that the proposed method is suitable for semi-automatic image classification purposes with satisfactory results. The map generated presented 7 land cover classes, which corresponded to the samples acquired. Through means of cross-validation the map exhibited a kappa accuracy of 0.73. The random forest classification model displayed good generalization power by being trained with samples of only one of the images (2016-10-30) and correctly detecting sand pit areas in the independent image (2016-05-03). Therefore, the development of an automated alert system for detecting sand pits in Amazonia is realistic and could provide spatial information for environmental agencies to regulate this activity. Future analysis will encompass the evaluation of field data and radiometric characteristics for impact assessment.

Keywords: Remote sensing, Sentinel-2, forest degradation, Amazonia, Multiresolution Segmentation, Random Forest, Classification.

1. Introduction

Due to its biodiversity, the Amazonian Forest composes one of the most important ecosystems on Earth (FEARNSIDE, 2013). Acre state is located in the lowlands of the southwestern Amazon, within a region considered with varied vegetal structure, such as: ombrophilous forest, palm trees, bamboos and campinaranas (GUIMARÃES; BUENO, 2016; SILVEIRA, 2003) being considered one of the most vulnerable and fragile Amazonian ecosystems (DALY et al., 2016). The soils of the region present as main characteristic sandy structure, nutrient poor and hydromorphic, leading to a greater time of regeneration after deforestation. Currently, among the most exploited natural resources for the region is the

extraction of sand (PAULA, 2011), which is often done in an irregular and undocumented way (RODRIGUES, 2007).

When exploited with a precarious methodology, the sand extraction activity is considered a harmful practice to the environment (SILVA; SIMI; RUDORFF, 2011). In the productive process, sand extracted from the stream bank is stored in fields located on its margins. The opening of these fields results in the complete elimination of ciliary vegetation. The subsequent removal of the sand with machines (loaders) causes intense degradation of the soil inside the fields due to the removal of the surface and sub-surface layers of the soil (JUNIOR et al., 2012).

When the sand mined in the river is scarce, the extraction dredges are moved to new locations, resulting in the abandonment of the courts. Devoid of vegetation cover and deeply altered soil, the recovery of the vegetation inside the fields is very impacted, damaging the recovery of the riparian forest that protects the banks of the river against the erosive process that causes the sedimentation of its bank.

With this environmental degradation process in course, remote sensing (RS) comes into play. The orbital RS is a key tool for the mapping and updating of critical areas, offering systematic coverage and high geometrical quality. In addition to these qualities resulting from the imaging in several multispectral channels, it is possible to get high spatial resolution images of those damaged areas. These tasks can be fulfilled with sensors like Sentinel-2.

To achieve frequent revisits and high mission availability, two identical Sentinel-2 satellites (Sentinel-2A and Sentinel-2B) are planned to operate simultaneously. The launch of the first satellite, Sentinel-2A, occurred on 23 June 2015 at 01:52 UTC on a Vega launch vehicle (INSIDER, 2015) Sentinel-2B will be launched in April 2017. However, there is still a spatial resolution problem regarding the Sentinel bands, they are not all in the same resolution (Table 1) and the sensor lacks a panchromatic band, which usually has a higher spatial resolution than the others.

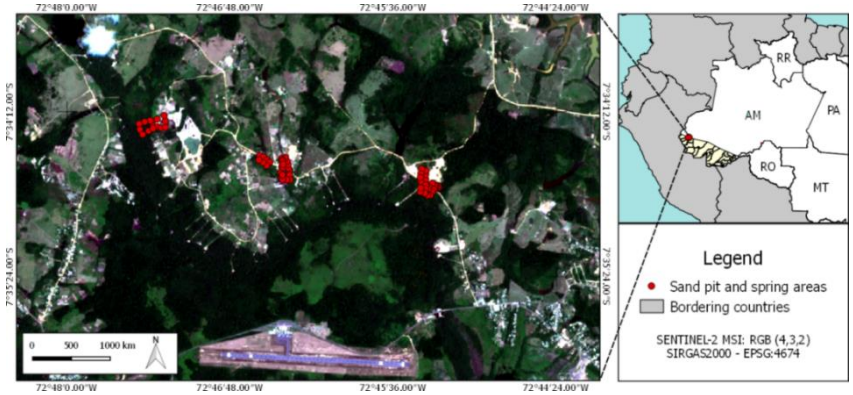
In order to address this need, the present paper proposes two objectives: (i) To evaluate the detection of sand pits in Sentinel-2 images, which are not large areas in extent but represent a severe harm to environment, and (ii) Asses the generalization power of the random forest (RF) model in one independent images of the same area.

2. Methods

2.1 Study site

The study area is situated in the municipality of Cruzeiro do Sul, state of Acre (Figure 1a). The terrain is formed by a series of hills and surrounded predominantly by Amazonian tropical dense vegetation. Cruzeiro do Sul presents a total area of 7,924km² (ACRE, 2013). It is located in the northwest region of the state of Acre, on the left bank of the Juruá River, 648 km by land from Rio Branco state capital connected by the BR-364 highway. The predominant type of soil is podzolic, red and yellow, not possessing rocky ground (DALY et al., 2016).

a)



b)

Figure 1.a) Location of the study area. The red dots, represents the three sand pits and water springs evaluated in this study; b) Google Street Maps panoramic view of the sand pit extraction site in Cruzeiro do Sul (Acre, Brazil).

2.2 Data

Two Sentinel-2A image products where used. The images were in processing level 1C, which accounts for top of atmosphere (TOA) reflectance. The data was downloaded from the European Space Agency (ESA) website <https://scihub.copernicus.eu/dhus/>. The product used is called S2A Level-1C which accounts for radiometric and geometric corrections (including orthorectification and spatial registration), and the quadrant T18MYS were selected. Two dates of imageries were acquired. The main image for developing the methodological procedure was acquired in 2016-10-30. The second image, acquired on 2016-05-03, was used as an independent dataset for testing the methodological generalization. The Sentinel-2A is equipped with the MultiSpectral Instrument (MSI) (Table 1).

Table 1.Spectral bands for the SENTINEL-2 Multispectral Instrument (MSI).

Band	Name	Central wavelength (nm)	Spatial resolution (m)
1	**Aerossol**	**443**	**60**
2	Blue	490	10
3	Green	560	10
4	Red	665	10
5	Red Edge 1	705	20
6	Red Edge 2	740	20
7	Red Edge 3	783	20
8	NIR	842	10
8a	Water-Vapor	865	20
9	**Cirrus**	**945**	**60**
10	**SWIR 1**	**1380**	**60**
11	SWIR 2	1610	20
12	Red Edge 4	2190	20

2.3 Data processing

The methodological procedure in this work can be divided in three main phases, presented below. Further details are presented in the following sections.

(i) the downloading of two satellite image data and pre-processing the images with the Science Toolbox Exploitation Platform (STEP) toolbox, which accounts for the atmospheric correction of the LV-1C Sentinel data.

(ii) the resulting pre-processed image data from phase (i) is then segmented in the eCognition platform by using its multiresolution segmentation algorithm. Posterior training samples were collected in this segmented image. The segmented image along with all the samples goes to final stage (iii).

(iii) in this phase, the segmented image and training samples are exported as ESRI shapefile and imported to the software [R] using a script. The script make a copy of the database files of

the original shapefile to a comma separated file that will be used by the Waikato Environment for Knowledge Analysis – WEKA (HALL et al., 2009) platform in the data-mining and further classification process. This script is freely available at <https://gist.github.com/davidguima/058e069a0d2d2573d296085339debd78>.

2.3.1 Multiresolution segmentation

It is not from nowadays that the segmentation is the object of study in the field of digital image processing (ROBERTS, 1963). Haralick and Shapiro (1985) define the segmentation process as the division of an image into several parts, being conditioned for this division that these parts share homogeneous properties between the elements that integrate it. These elements can be pixels or even other segments, and properties that must share similarities vary depending on the application one expects from the segmented image.

Within the RS techniques, the segmentation process precedes the classification stages that are based on regions, and among the segmentation algorithms available, the segmentation algorithm developed by Baatz and Schape (2000) stands out. This segmentation algorithm is called by the authors as polyvalent in view of its multiplicity of applications, and it is adapted to the scale of the objects presented in the images according to the parameters entered by the user, which is known worldwide as multiresolution or multiscale targeting.

The segmentation parameters for this research were defined as follows: *Shape* = 0, *Compactness* = 0 and *Scale parameter* = 70. Both images where individually segmented with the same parameters, which also account for *Image Layer Weights* = 0 for bands 1,6,7,9,10,11,12,8A; 1 for bands 2,3,5,8 and 10 for band 4. Alongside the multiresolution segmentation, the Spectral Difference algorithm was also used to merge neighboring objects according to their mean DN intensity values. The parameters for this algorithm are *Maximum Spectral Difference* = 90; *Image Layer Weights*= 0 for bands 1,9,10; 1 for bands 2,3,6,7,8,11,12,13; 2 for band 5 and 10 for band 4.

2.3.2 Data-mining algorithms and Random Forests

Objects present in the images (or segments) carry along their characteristics like shape, spectrum, hierarchical information and statistics. These properties are used as information source to define separation thresholds by the mining algorithm, which will further define the inclusion or exclusion of such parameters by the image classification algorithm. This study has made exclusive use of the *mean layer value* as a way of representing the spectral information inside the pixels of the image. The method for mining the data and classification was the Random Forest algorithm proposed by Breiman (2001) and further implemented in the WEKA platform (HALL et al., 2009).

2.3.3 Validation

The evaluation of the generated thematic map was performed by using the Kappa index and a k-fold cross validation over the 73 samples that were acquired in the segmented image, considered as ground truth data. The Kappa coefficient of agreement is constructed from an error matrix, in which errors of omission are expressed, that is, samples that were not classified according to the reference classes, and the commission errors corresponding to samples of incorrectly referenced as belonging to other classes (CONGALTON; GREEN, 2009). The Equation 1 gives the Kappa index.

$$K = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k}(n_{i+}n_{+i})}{n^2 - \sum_{i=1}^{k}(n_{i+}n_{+i})}, \quad (1)$$

Where n_{ii} – total number of samples correctly classified from class k; n_{i+} – total number of classified samples from class k; n_{+i} – total number of collected samples from class k; and n – total number of samples.

In addition to the Kappa index, one of the authors (LR) has performed field work in the impacted areas. The map was visually inspected by her and evaluated against more than 80 field Global Positioning System (GPS) data points (Table 2).

Table 2. Strata of 10 out of the 81 GPS points collected *in situ* for Spring and Igarapé areas.

Spring	A	B	C	D	-	-
South	7º 34'44,93"	7º 34'43,23"	7º 34'36,64"	7º 34'16,14"	-	-
West	72º 45'19,27"	72º 45'23,40"	72º 46'22,34"	72º 47'22,36"	-	-
Igarapé	A	B	C	D	E	F
South	7º 34'45,06"	7º 34'41,76"	7º 34'36,16"	7º 34'36,35"	7º 34'31,10"	7º 34'15,86"
West	72º 45'19,84"	72º 45'24,11"	72º 46'25,26"	72º 46'24,54"	72º 46'35,04"	72º 47'23,36"

3. Results

Preliminary results on the image segmentation exhibits a spatial resolution gain from 60 to 10 meters per pixel (Figure 3a), by using the mean value of the digital numbers inside the segments of the multiresolution algorithm output (Figure 3b). This segmentation made possible to improve the original coarse image (Figure 3c) to be resampled to the same resolution (10 m) as the fine spatial resolution Red, Green, Blue and NIR bands.

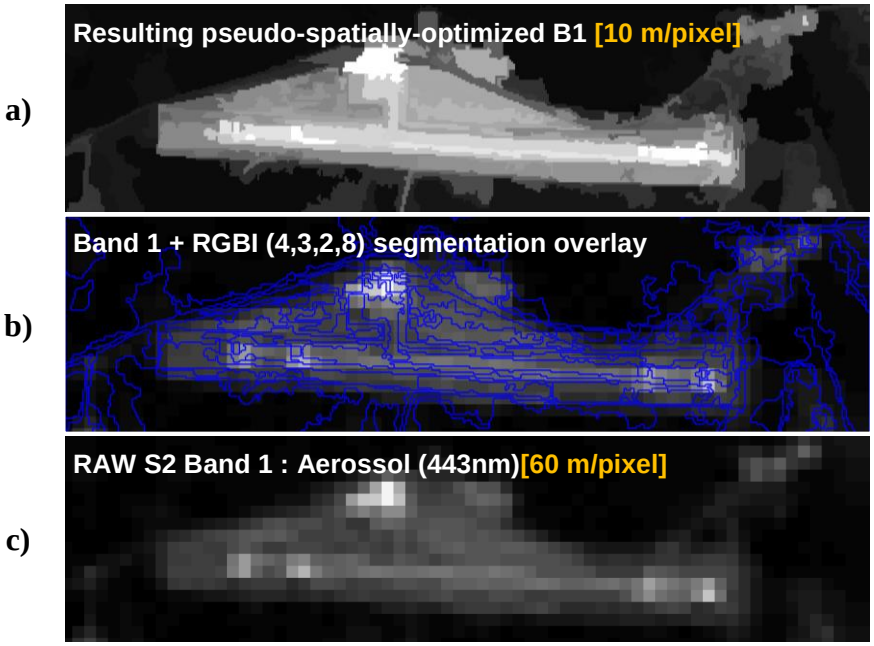


Figure 3. **a)** Resulting further-detailed image rescaled from 60m to 10m.; **b)** Multiresolution segmentation applied in the 10m bands overlaid the in the coarse original S2-A band 1; And **c)** Original Sentinel-2 Band 1 image with coarse (60m) spatial resolution.

By making use of the new generated mean layers for bands 1 to 8A, we classified the image, generating as a result a thematic map for each one of the images of the study site (Figure 4). The definition of the selected classes are inherited from the original TerraClass data (ALMEIDA et al., 2016) with minor adaptation of the typologies for the purpose of including the sand pit areas. Some classified class confusion could be observed with the inclusion of the clouds in the "*not observed"* class and some minor confusions of the land cover classes "*Bare soil"* with "*Urban"* class, which was already expected assuming the influence of the Red wavelength (B4 - 665nm) on those classes and thus its spectral similarity.

a)

b)

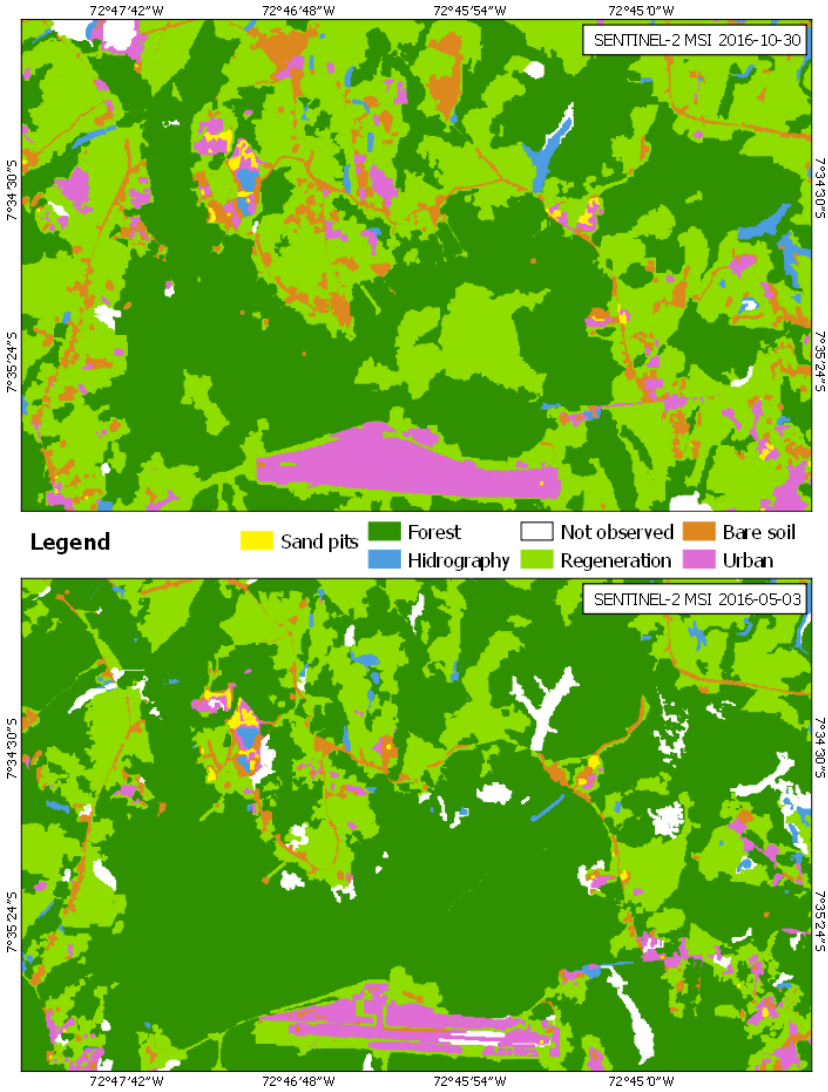


Figure 4. Final results of the Random Forest classification in Waikato Environment for Knowledge Analysis making use of the Cirrus band data in the model generated from the **a)** 2016-10-30 onto the **b)** 2016-05-03 image.

The cross validation over the samples collected from the image 2016-10-30 displayed a Kappa coefficient of 0.73 and is presented in the Table 4. Furthermore, the percentage of land area taken by each class is also estimated and presented in subsequent Table 5.

Table 4. Random Forest 10 folds cross-validation output statistics and confusion matrix over the classifier model, trained with the 2016-10-30 image sample segments.

A	B	C	D	E	F	G	<- Classified as
9	0	0	0	1	1	0	A = Not observed
0	9	0	0	0	0	0	B = Forest
0	0	5	1	0	0	4	C = Bare soil
0	0	2	9	0	0	0	D = Regeneration
2	0	0	0	6	0	0	E = Hidrography
0	0	0	0	0	8	2	F = Sand pits
0	0	1	2	0	1	10	G = Urban
0.82	1.00	0.50	0.82	0.75	0.80	0.82	Weighted avg.
TP-Rate							**0.767**

Correctly Classified Instances	56 (76.7123 %)
Incorrectly Classified Instances	17 (23.2877 %)
Kappa statistic	**0.7264**
Mean absolute error	0.11
Root mean squared error	0.2318
Relative absolute error	45.04%
Coverage of cases (0.95 level)	98.63%
Mean rel. region size (0.95 level)	36.99%
Root relative squared error	66.29%
Total Number of Instances	73

In order to visually assess the spectral characteristics of the thematic classes, the Figure 5 displays the normalized reflectance of six of them. The class "Not Observed" was not plotted as it is just an abstraction of all the unwanted targets of the scene (clouds and shadow).

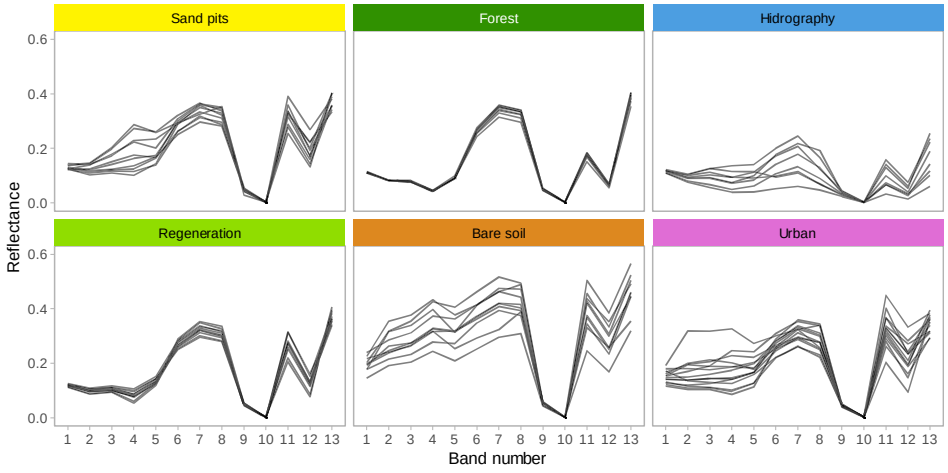


Figure 5. Normalized reflectance spectra of the six thematic classes, besides the "Not observed" class, which was omitted for having mixed spectral behavior resulting from clouds and shadows in the same sample group.

Table 5. Percentage (%) of area taken by each thematic class in the two Sentinel-2 image dates.

CLASS	05-03-2016	10-30-2016
Sand pits	0.25	0.29
Forest	62.65	42.27
Hydrography	1.26	1.62
Not observed	2.93	1.08
Regeneration	26.94	43.63
Bare soil	2.53	5.38
Urban	3.44	5.74
TOTAL (%)	100.00	100.00

4. Considerations

Regarding the Sentinel-2 sensor, and in the absence of a panchromatic band, the proposed methodology displayed satisfactory results by optimizing the spatial resolution from 60 to 10 meters in the Aerosol(1), Water-Vapor (9) and Cirrus(10) bands. It is, however, worth mentioning that the segmentation process was impacted by the direct spatial influence of the Red (4), Green (3), Blue (2) and NIR (8) bands of this sensor as an inherited property of the multiresolution algorithm. As a result, any possible radiometrically spurious pixels (i.e. speckle noise, sun glint and/or bad digital number-DN values) in these bands possibly directly influenced the shape of the resulting optimized band. Nonetheless, the information in the resulting image was derived from the same band in which the radiometric statistical mean was extracted. Thus, it is possible to assume that only the spatial information of the resulting pseudo-spatially optimized band was compromised by any lacking radiometric information from those bands.

Furthermore, the image segmentation extracted the mean value of all the pixels inside a given image segment, which means that this method causes a generalization of the spectral information intra-segment. That generalization is not recommended for spectral image analysis purposes (i.e. atmospheric correction), but only for object-based image analysis (OBIA) and other methods that make extensive use of the spatial resolution of the sensor imageries.

As of the semi-automatic image classification results, the development of an automated alert system for detecting sand pits in Amazonia is realistic and could provide spatial information for environmental agencies to regulate this activity. Future analysis will encompass the statistical validation of field data and radiometric characteristics for impact assessment. The final results for this working in progress methodology where overall satisfactory and could also be implemented in similar scenarios where panchromatic sensor data is non-existent.

Acknowledgments

The authors would like to thank the National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN) for the first author scholarship (Process No:312500/2016-5), the National Institute for Space Research (INPE) and the Tropical Ecosystems and Environmental Science Laboratory (TREES) for providing the infrastructure to develop this research. MM thanks the Brazilian National Council for Scientific and Technological Development (CNPq) for her scholarship (process number 313065/2015-2). SS would like to thank the Research Council of Norway (project number 230860 – https://www.nhh.no/en/research-centres/tropical-deforestation/) for the financial support.

References

ACRE, G. DO E. DO. **Acre em números 2013**. Disponível em: <http://www.ac.gov.br/wps/portal/acre/Acre/estado-acre/sobre-o-acre>. Acesso em: 7 nov. 2016.

ALMEIDA, C. A. DE et al. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. **Acta Amazonica**, v. 46, n. 3, p. 291–302, 2016.

BAATZ, M.; SCHAPE, A. Multiresolution Segmentation : an optimization approach for high quality multi-scale image segmentation. **Journal of Photogrammetry and Remote Sensing**, v. 58, p. 12–23, 2000.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CONGALTON, R. G.; GREEN, K. **Assessing the Accuracy of Remotely Sensed Data: Principles and Practices**. [s.l.] CRC Press/Taylor & Francis, 2009.

DALY, D. C. et al. The White-sand Vegetation of Acre, Brazil. **Biotropica**, v. 48, n. 1, p. 81–89, jan. 2016.

FEARNSIDE, P. M. Climate Change as a Threat to Brazil's Amazon Forest. **International Journal of Social Ecology and Sustainable Development**, v. 4, n. 3, p. 1–12, 2013.

GUIMARÃES, F. S.; BUENO, G. T. As campinas e campinaranas amazônicas / The amazonian campinas and campinaranas. **Caderno de Geografia**, v. 26, n. 45, p. 113, 30 dez. 2016.

HALL, M. et al. **The WEKA data mining softwareACM SIGKDD Explorations Newsletter**, 2009. Disponível em: <http://portal.acm.org/citation.cfm?doid=1656274.1656278\npapers2://publication/doi/10.1145/1656274.1656278>

HARALICK, R. M.; SHAPIRO, L. G. Image segmentation techniques. **Computer Vision, Graphics, and Image Processing**, v. 29, n. 1, p. 100–132, jan. 1985.

INSIDER, S. **Arianespace successfully launches Europe's Sentinel-2A Earth observation satellite**. Disponível em: <http://www.spaceflightinsider.com/missions/earth-science/arianespace-successfully-launches-europes-sentinel-2a-earth-observation-satellite/>. Acesso em: 17 ago. 2016.

JUNIOR, N. L. L. et al. IMPACTO DA MINERAÇÃO DE AREIA NAS CARACTERÍSTICAS FLORÍSTICAS DE UM FRAGMENTO DE MATA CILIAR EM REGENERAÇÃO NAS MARGENS DO RIO ACRE, EM RIO BRANCO-AC. **64ª Reunião Anual da SBPC**, 2012.

PAULA, M. A. S. DE. **Avaliação dos estudos de impactos ambientais realizados no Acre, no período de 1989 a 2005**. [s.l.] Universidade de Brasilia, 2011.

ROBERTS, L. G. **Machine perception of three-dimensional solids**. [s.l.] Massachusetts Institute of Technology, 1963.

RODRIGUES, E. **Mineração e degradação ambiental no Acre**. Disponível em: <http://www.andiroba.org.br/artigos/?post_id=1703>.

SILVA, G. B. S. DA; SIMI, R.; RUDORFF, B. F. T. Monitoramento da extração de areia nos municípios não pertencentes ao Zoneamento Ambiental Minerário do trecho paulista da várzea do rio Paraíba do Sul. **Anais XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR**, p. 4877–4884, 2011.

SILVEIRA, M. **VEGETAÇÃO E FLORA DAS CAMPINARANAS DO SUDOESTE AMAZÔNICO (JU-008)**. [s.l.] Universidade Federal do Acre, 2003.