Revista Brasileira de Cartografia
RBC | Rev. Bras. Cartografia

DOI: 10.14393/rbcv70n5-45180

Information mining for automatic search in remote sensing image catalogs

Mineração de informações para busca automática em catálogos de imagens de sensoriamento remoto

Mikhaela Aloísia Jéssie Santos Pletsch[1]
Thales Sehn Körting[1]

Recebido em abril de 2018.
Aprovado em dezembro de 2018.

ABSTRACT
The Earth Observation database is almost on the scale of Zettabyte (10^{21} Bytes). Produced at a rapid rate, those data also present great diversity, due to the range of sensor types. In such a manner, this kind of data is also classified as Big Data, and present opportunities, such as the possibility to analyze and integrate different data, as well as challenges, mainly regarding storing and processing steps in order to be available to users. The distribution of this database is normally through catalogues, which searching criteria are limited to traditional metadata, as acquisition date, sensor characteristics and geographical localization. Thus, there is a demand for a tool that enables users to search for images based on phenomena in lieu of date or location in a data fusion perspective. In this manner, this work resulted in a Remote Sensing Image Information Mining (ReSIIM) prototype able to make smart searches in big databases based on well-known and basic targets found in Remote Sensing imagery: *cloud, cloud shadow, clear land* (land area), *water, forest, bare soil, built-up* and *burned area*. For that, the aforementioned targets metadata are extract and stored in databases, enabling to refine and boost searches. Besides the spatialization and discussion of the aforementioned targets along Brazil, ReSIIM was assessed according to its accuracy to identify adequate images in databases for Linear Spectral Mixture Model (LSMM) application, which is one of the main methods available to detect burned areas in the Brazilian Amazon. With the adequate search criteria, ReSIIM was able to retrieve all adequate images, indicating its high potential for this and further applications.

KEYWORDS: Image Retrieval. Burned Area Detection. Amazonia. Satellite imagery. Data Mining.

[1] Instituto Nacional de Pesquisas Espaciais - INPE. Av. dos Astronautas, 1.758 - Jardim da Granja, São José dos Campos - SP, Brasil, 12227-010. E-mail: mikhaela.pletsch; thales.korting@inpe.br

RESUMO
A base de dados de observação da Terra está se aproximando da escala de zettabyte (10^{21} Bytes). Produzidos em alta velocidade, estes dados apresentam também grande variedade, devido a fatores como a diversidade dos sensores existentes. Frente a estas características, são enquadrados como Big Data, e apresentam tanto oportunidades, como a possibilidade de analisar e integrar um maior conjunto de dados, quanto desafios, principalmente no armazenamento e processamento para que possam ser disponibilizados aos usuários. A distribuição desta base é geralmente realizada por meio de catálogos, cujos parâmetros de buscas ainda são limitados aos metadados tradicionais de imagens de satélite, como data de aquisição, característica do sensor e localização geográfica da área imageada. Assim, ainda existe uma demanda em oferecer aos usuários parâmetros com base no conteúdo das imagens em uma perspectiva de fusão de dados, permitindo buscas mais refinadas para dar suporte a estudos específicos e possíveis tomadas de decisão. Nesse contexto, o presente trabalho resultou em um protótipo de mineração de informações de imagens de Sensoriamento Remoto (*Remote Sensing Image Information Mining* - ReSIIM), capaz de realizar buscas inteligentes em grandes catálogos de imagens com base nos alvos de *nuvem, sombra de nuvem, área continental, água, floresta, solo exposto, área construída* e *área queimada*. Para isto, os metadados dos alvos ora mencionados são extraídos e armazenados em bancos, permitindo refinar e facilitar pesquisas. Além da espacialização e discussão dos alvos mencionados ao longo do Brasil, o ReSIIM foi avaliado de acordo com o seu potencial de identificar imagens adequadas em bancos de dados para a aplicação do Modelo Linear de Mistura Espectral (MLME), o qual é um dos métodos disponíveis para detectar áreas queimadas na Amazônia Brasileira. Com o critério de busca apropriado, o ReSIIM foi capaz de retornar todas as imagens adequadas, indicando o alto potencial da ferramenta para esta e outras aplicações.

PALAVRAS-CHAVE: Recuperação de Imagens. Detecção de Áreas Queimadas. Amazônia. Imagens de Satélite. Mineração de Dados.

* * *

Introduction

Analyzing large areas of the Earth's surface in a short time is possible using Remote Sensing (RS) tools. However, describing and finding information, as well as improving data management, analysis and cataloging in such a great amount of data are some of the main challenges (LI et al., 2016). According to Quartulli and Olaizola (2013), Earth Observation (EO) data are gradually achieving the scale of a zettabyte (1 ZB = 10^{21} Bytes). Produced in a high velocity, EO data come from several remote sensors in different data types, resolutions and scales (DATCU and SEIDEL, 2002,

2003; LI et al., 2016). In addition, the aspect of the atmosphere in the moment of the image acquisition as well as the ways of capturing the spectral interaction between the electromagnetic radiation and objects characterize its dense diversity (KÖRTING et al., 2016). Considering the data deluge, production velocity and high diversity (LANEY, 2001), EO data is also coined as Big Data (KÖRTING et al., 2016). Its management is a conundrum once it presents not only opportunities but also drawbacks such as comprehending complex spatial phenomena in a fast and accurate way (CHI et al., 2015; LI et al., 2016). In this manner, continuously improvements and updated RS image query tools are necessary.

Aiming to organize such amount of data, RS image catalogs were developed by institutions in order to manage and distribute EO data. Many efforts currently aim to develop efficient, easily-accessible and well-sourced catalogs. Some examples are the United States Geological Survey (USGS - https://data.usgs.gov/datacatalog/), the European Space Agency (ESA - *https://earth.esa.int/web/guest/data-access/catalogue-access*) and the Brazilian National Institute for Space Research (INPE - *http://www.dgi.inpe.br/catalogo/*). Through such catalogs, it is possible to search for images according to user's specifications, such as location and date (DATCU et al., 2000; LI and NARAYANAN, 2006; STEPINSKI et al., 2014).

A great amount of satellite images are freely available, yet they are underexploited once a great volume has never been analyzed (QUARTULLI and OLAIZOLA, 2013) or scientifically evaluated in a context or connected through time. This issue may be overcome with more accurate images search tools, which are not available even in the most modern satellite image catalogs, indicating that smart tools are scarce (STEPINSKI et al., 2014) in order to improve geospatial data handling (QUARTULLI and OLAIZOLA, 2013). In such a way, an alternative is to provide approaches that extract and generate metadata in satellite images about well-known targets, such as water and forest, besides the commonly available cloud masks.

1. Image Retrieval

Firstly, image retrieval was based on text, however, some targets are hardly described by words and it also requires manual work. As a consequence of the aforementioned gaps, Content-Based Image Retrieval (CBIR) was developed (DATCU and SEIDEL, 2003; EAKINS and GRAHAM, 1999; GOODRUM, 2000; LONG et al., 2003). In CBIR systems, images are retrieved based on its visual similarities considering a given group of characteristics. Nonetheless, taking into account the complexity identified in such volume of data, advances focused on the user were lacking (DATCU and SEIDEL, 2003). In this context, in the end of 90's the concept of *Knowledge-Discovery in Databases* (KDD) got incorporated by CBIR, engendering the so called *Image Information Mining* (IIM). Although between CBIR and IIM there is a subtle overlap (LI and NARAYANAN, 2006), IIM aims to extract the main implicit information in databases (DATCU et al., 2000) through different methods and approaches, which advances may offer an unique opportunity to discover the content present in each image (SHYU et al., 2007).

IIM is equivalent to the communication step between a database and users, whose accuracy varies according to the approaches (DATCU and SEIDEL, 2003) used to identify targets along scenes.

1.1 Remote Sensing Image Metadata

According to Guptill (1999), metadata provides detailed attributes and characteristics description of a given element, being *the data about data*. As such, it is unavoidable once it is essential in a filing process (EAKINS and GRAHAM, 1999). Based on RS image metadata, different search criteria can be developed in catalogs. Roughly, a catalog refers to a description list of items found in a collection (FRANK, 1994). A data catalog is thus a collection of metadata records, which is associated with search tools and data management. Therefore, an image catalog facilitates the operations of

searching, sharing and processing data to users, and may support decision makers (GUPTILL, 1999).

RS image metadata can be identified through some approaches, such as the Fmask algorithm, first developed by Zhu and Woodcock (2012) and improved by Zhu et al. (2015). Based on these works, Flood and Gillingham (2018) developed a set of command line utilities in a Python module. The output of the algorithm is a single thematic raster with up to 6 different values [0, 5], representing: *null value*, *cloud*, *cloud shadow*, *clear land*, *snow*, and *water*. *Clear land* refers to any data that is neither *cloud*, *cloud shadow*, *snow* nor *water* and even though presents some valid information.

Regarding water targets, Namikawa et al. (2016) extracted 5-meter spatial resolution masks from Brazilian water bodies using RapidEye images with an automated methodology. Although water bodies are usually identified due to its low reflectance, in the real world several parameters may interfere their detection, such as suspended solids and water depth. With this in mind, the methodology was based on the color transformation from Red-Green-Blue (RGB) to Hue-Saturation-Value (HSV) and the minimum radiance from all the bands. For that, some factors were considered, as the differences in illumination and scattering throughout more than 15,000 RapidEye scenes. As a result, it was possible to classify seven classes of water in agreement with the confidence of the classified pixels ranging from 1 to 7, according to the presence of water, where 1 is more reliable and 7, less reliable. Furthermore, the degree of persistence is also available. According to the authors, although this methodology is considered simple, it is accurate to detect water bodies. Namikawa and Castejon (2017) identified some issues which may interfere the results nonetheless, such as cloud noise, shadow in urban areas and specular reflection of sunlight.

Different channel combinations of multispectral RS images can be also used to enhance complex phenomena identification (BANNARI et al., 1995; KEY and BENSON, 2006), such as vegetation indices that enhance the contribution of vegetation properties (HUETE et al., 2002). For instance, the

Enhanced Vegetation Index (EVI), which is more sensitive to topographic conditions, and the Normalized Difference Vegetation Index (NDVI), more sensitive to chlorophyll (GAO et al., 2000; MATSUSHITA et al., 2007). NDVI was developed by Rouse et al. (1974), and it ranges from [-1.0, +1.0]. The definition of NDVI thresholds for mapping vegetation is controversial, and it also varies according to the region. In a general way, NDVI values between 0.0 and 0.1 refer to rocks and bare soil. Values greater than 0.1 indicate a gradual increase in *greenness* and intensity of vegetation (NOAA, 2017), although some authors use the threshold of 0.2 for bare soil and vegetation (JIN et al., 2014; LIANG et al., 2013), and the transition zone is considered from 0.1 to 0.2 (LIANG et al., 2013).

Other indices commonly used, but with controversial thresholds are: Bare Soil Index (BSI) (RIKIMARU et al., 2002) for bare soil identification, the Normalized Difference Built-up Index (NDBI) for built-up area detection (ZHA; GAO; NI, 2003), and the Burned Area Index (BAI) (MARTÍN ISABEL, 1998), and the Normalized Burn Ratio (NBR) (GARCÍA and CASELLES, 1991; KEY and BENSON, 2006) for burned areas detection.

1.2 Burned Areas in Brazilian Amazon and the use of a Linear Spectral Mixture Model

Several rainforests worldwide are located in underdeveloped countries. In such a way, resources to protect and preserve this kind of environment are typically scarce (FAO and ITTO, 2011). Although rainforests play an important role in climate regulation, they face countless threats. Among them, in the Brazilian Amazon, deforestation is often the first one to be pointed, yet during a fire, the main gas emitted is carbon dioxide, which is also the primary Greenhouse effect gas (ANDERSON et al., 2005b; LASHOF, 1991). According to Aragão and Shimabukuro (2010), drought years followed by fires release as much carbon (C) as deforestation processes. The negative effects of fires thus extend beyond damage to a single swath of trees. They

influence global climate changes once surface radiative changes have occurred (ANDERSON et al., 2015; PADILLA et al., 2017; SHIMABUKURO et al., 2009, 2015). Therefore, comprehending, managing and avoiding fires in the Amazon region could meet the international demand for C emission reductions. These fires, normally induced by humans, are often applied to facilitate land use and land cover (LULC) changes. Where before there was a natural habitat supporting a wide variety of life forms, agricultural areas and pastures arise (ARAGÃO et al., 2016; SHIMABUKURO et al., 2015).

Currently, burned forest research in the Brazilian Amazon is commonly performed based on the Linear Spectral Mixture Model (LSMM) (ANDERE et al., 2015; ANDERSON et al., 2005a, 2015; CARDOZO et al., 2013; SHIMABUKURO et al., 2009), once it presents the most accurate results up to now in this biome. Even for more accurate spatial resolutions, satellite data presents a *mixture problem* (ANDERSON et al., 2005a; SHIMABUKURO et al., 2009), since a pixel represents the average spectral response from all the elements located in that pixel. In this perspective, LSMM was developed aiming to depict subpixel heterogeneity (SHIMABUKURO and SMITH, 1991). In this model, some pure pixels called *endmembers* are selected by a domain specialist, deriving shade fraction images for burned area detection. Similar spectral responses may interfere with the results though (ANDERE et al., 2015; BASTARRIKA et al., 2011; CHUVIECO and CONGALTON, 1988). For instance, burned areas exhibit low reflectance, as well as water and cloud shadows. Moreover, cloud coverage and fire smoke may also omit pixels with spectral response affected by fire (ARAGÃO et al., 2016), whilst clouds and their shadows influence negatively many uses of EO data, such as inaccurate atmospheric correction and land cover classification (ZHU and WOODCOCK, 2012). In this context, applying LSMM methodology requires a prior manual step: identifying appropriate images, free of clouds and with the potential to show burned areas.

In this context, the main contribution of this paper is the presentation of Remote Sensing Image Information Mining (hereafter ReSIIM), which is an

innovative approach for remote sensing image retrieval able to make smart searches in big databases based on well-known and basic targets found in Remote Sensing imagery. Furthermore, we also present the results of ReSIIM along Brazil based on the presence of *cloud*, *cloud shadow*, *water*, *clear land*, *forest*, *bare soil*, *built-up* and *burned areas*, which were developed based on Fmask and spectral indices rules.

This paper also presents some tests to verify ReSIIM's potential as a support tool to retrieve from a database only appropriate images for LSMM application. ReSIIM is, nonetheless, versatile enough to be applied on any related RS approaches, and was idealized to be continuously improved according to demands.

This paper is an extended version of Pletsch and Körting (2017), presented in XVII Brazilian Symposium on GeoInformatics (GEOINFO 2017).

2. Methodology

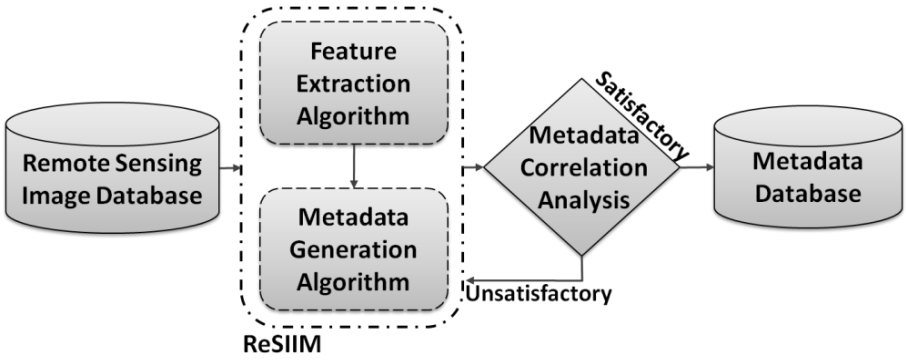
In this section, we present the developed methodology (Figure 1). The Remote Sensing image database is composed of Landsat 5 and 8 imagery. ReSIIM is organized in two main steps, feature extraction algorithm and metadata generation algorithm, which were developed through different approaches and rules (Section 2.1). In a decision process, an automatic correlation analysis is applied to validate the previous classification based on a given reference dataset (Section 2.2). If the result is unsatisfactory, it goes back to the ReSIIM phase in an attempt to generate more accurate metadata. Finally, the metadata is stored according to the results obtained by ReSIIM. No attention is drawn to searching techniques or data storage, once we focused to refine and boost searches through metadata.

With the obtained results, all data were spatialized and discussed in Section 3.1 (Figure 5-6) in order to analyze the potential of using the obtained results along Brazil. For that, Landsat scenes along Brazil during the Amazon dry season (between August and September, according to Aragão et

al. (2007)) from 2017 were used. With almost 13 million km² of data in 0.75TB, the whole processing step took about three days.

The potential of ReSIIM as a support tool for LSMM application was analyzed according to a set of tests. According to some searching criteria, we analyzed the potential of ReSIIM to retrieve only data suitable for LSMM application. Because indices present yet controversial thresholds, in this step, only targets derived from Fmask algorithm were analyzed. Results are presented in Section 3.2.

Figure 1 – Methodology Flowchart



Source: Elaborated by the authors.

2.1 Remote Sensing Image Information Mining (ReSIIM)

Our Remote Sensing Image Information Mining methodology aims to extract and generate metadata from satellite images through consolidated methods and indices found in the literature (Section 2.1.1). An example of ReSIIM results is found in Section 2.1.2.

2.1.1 Feature extraction and metadata generation algorithms

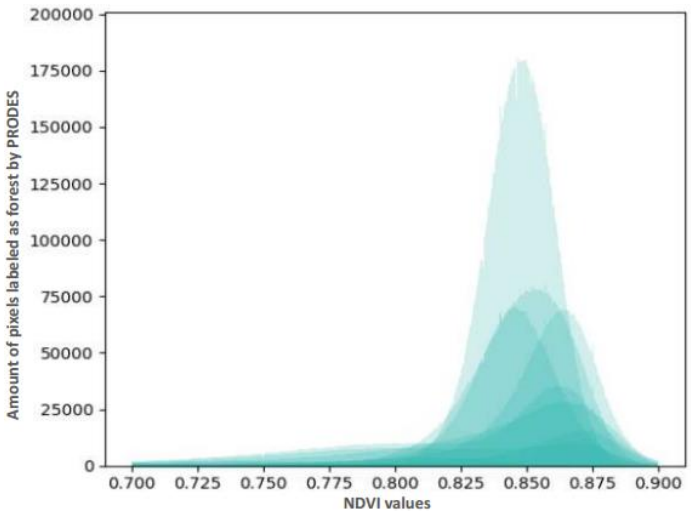
In this work, the first step involves the categorization of *cloud*, *cloud shadow*, *water* and *clear land* through Fmask algorithm (ZHU et al., 2015; ZHU and WOODCOCK, 2012). In such a way, Fmask algorithm targets with

valid data compose 100% of an image. Only for *clear land* areas, different indices were applied to *forest*, *bare soil*, *burned* and *built-up* area detection.

Considering the several meanings that the land cover *forest* carries and the different NDVI values found along such areas, we aimed to narrow the scope of NDVI based on Amazon forest. For that, we used forest masks classified by the Brazilian National Institute for Space Research (INPE) from the PRODES project, which monitors the Brazilian Amazon through satellites (INPE, 2017). After analyzing the NDVI values found in the pixels labeled as forest by PRODES (Figure 2), we identified that values of NDVI above 0.8 present a high correlation to dense vegetation. As such, we defined the rule NDVI ≥ 0.8 for *forest* target detection.

For *bare soil*, a combined rule was applied aiming to strict the results, as well as for *burned areas*, based on literature review and empirical analysis throughout Brazilian territory. The rule used for *built-up* areas aimed to highlight built-up targets (NDBI ≥ 0.0) at the same time that bare soil was not considered in this classification step (BSI ≤ 0.0), once NDBI is unable to separate urban from barren areas, due to the similar spectral response (ZHA et al., 2003). For *burned areas* rules, BAI and NBR were used based on empirical analyses. The rules used in ReSIIM are presented in Table 1.

Figure 2 – NDVI values frequency in areas identified as forest by PRODES in different masks - The darker, the more frequent the value is



Source: Elaborated by the authors.

Table 1 - Target rules used in ReSIIM.

Targets	Rules
Forest	NDVI ≥ 0.8
Bare Soil	0.0 ≤ NDVI ≤ 0.2; BSI > 0.0
Built-up	NDBI ≥ 0.0; BSI ≤ 0.0
Burned areas	BAI ≥ 250; NBR > 0.1

Source: Elaborated by the authors.

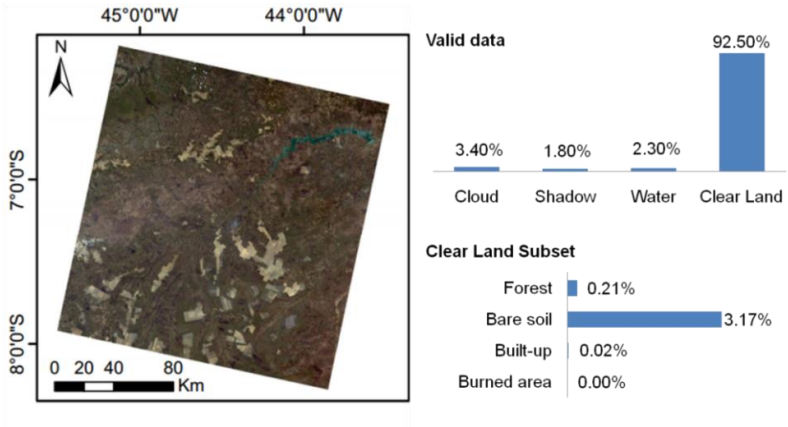
2.1.2 ReSIIM results example

An example of ReSIIM application on a RS scene is presented in Figure 3 (path/row 220/065 - 24.09.2013). With high percentage of *clear land*, it is possible to analyze in a more precise way possible targets that occur as a subset of this group.

2.2 Metadata Correlation Analysis

In this section, we explain the methods used to analyze the correlation between the generated metadata and real targets along a set of images. For that, random Landsat (L8) scenes along Legal Amazon were used (freely available at *https://earthexplorer.usgs.gov/*).

Figure 3 – Example of ReSIIM application on a Remote Sensing image and the generated metadata (scene 220/065 - color combination in R4G3B2 using L8 bands)



Source: Elaborated by the authors.

Due to the lack of official source data in Brazilian Amazon neither *cloud*, *cloud shadow*, *clear land*, *bare soil*, *built-up,* nor *burned areas* correlation analyses were possible. As highlighted by Liang et al. (2013) for the metadata *bare soil*, validation processes in this area of research are restricted due to limited data sources and methodologies.

2.2.1 Water

The correlation analysis of *water* was performed based on the reference dataset developed by Namikawa et al. (2016). This approach was taken into account, once it is more flexible when compared with a common threshold for a large amount of scenes (NAMIKAWA et al., 2016). The *water* reference data was filtered according to its high persistence along 4 years.

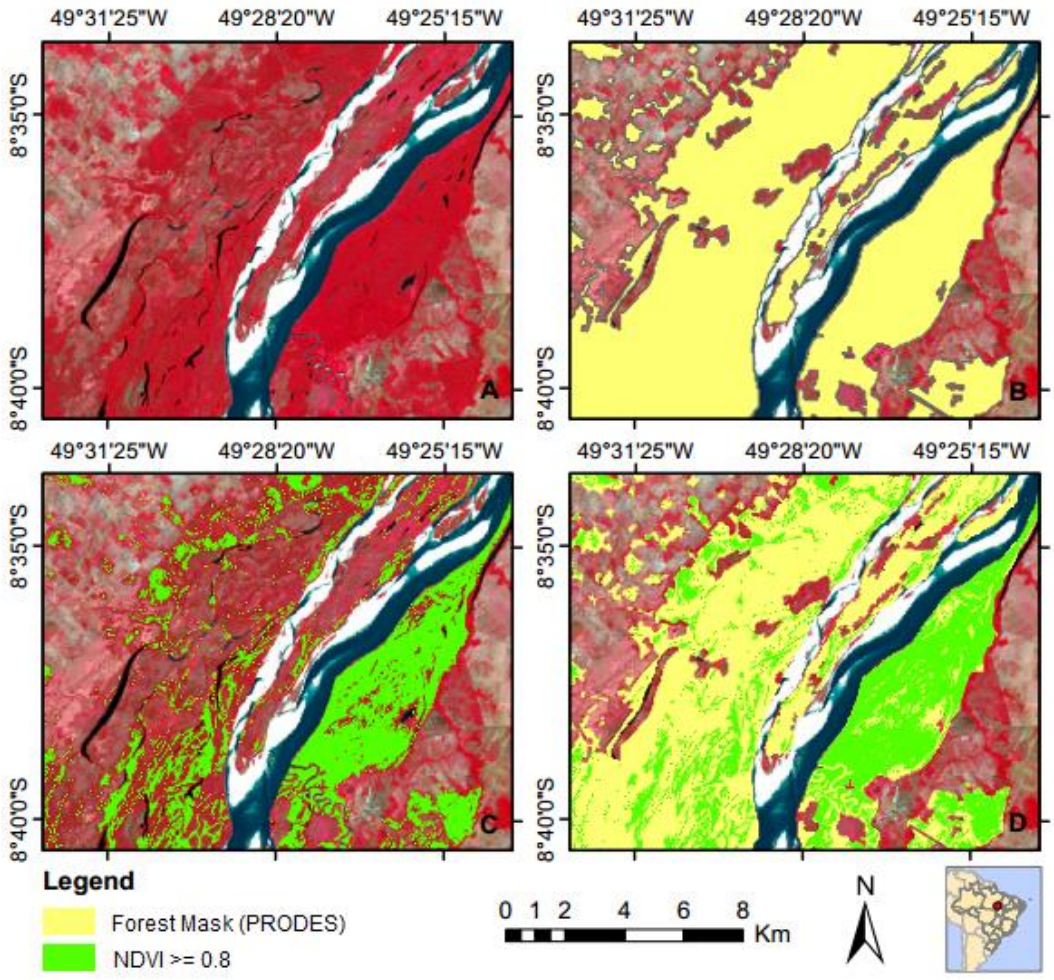
The correlation analysis was based on two main approaches: water correlation accuracy (WCA) and commission error (CE). WCA refers to the overlapped data between the target water generated in ReSIIM and the reference data, whilst CE refers to the data wrongly classified. For that, 10 random scenes were selected along Brazilian Amazon. WCA ranged from 76% to 99% of correct classification, and average WCA was 90%. In this manner, the misclassifications were not higher than 24%, and average CE was about 10%. Omission error was not taken into account, since the reference data's spatial resolution is 5m and the classified data is 30m (L8).

2.2.2 Forest

Our *forest* correlation analysis was based on two approaches: correlation accuracy (CA) and commission error (CE). CA refers to the overlapped data between the target *forest* generated in ReSIIM and the forest reference data from PRODES, whilst CE refers to the misclassified data. For this analysis 10 random scenes were selected from the Brazilian Amazon.

The results were satisfactory, considering that average CA was 93% and average CE was 7%. Nonetheless, it is remarkable that although CA was higher than 92% for each scene, just in a riparian vegetation scene, CA was 58% and CE, 42%. Figure 4 presents this special case in four steps: A - original data; B - forest reference (PRODES); C - detected forest (NDVI ≥ 0.8); and D - final composition of the three aforementioned layers. Along the bank of the river, the riparian forest presents NDVI values ranging mainly from 0.4 to 0.7. This outline suggests that the used threshold is not suitable for this kind of vegetation.

Figure 4 – Outlier identified in NDVI threshold for forest target. A - Original Scene (color combination in R5G4B2 using L8 bands); B - Forest Mask (PRODES); C - Detected Forest (NDVI ≥ 0.8); D - Input Image, Forest Mask, and Detected Forest



Source: Elaborated by the authors.

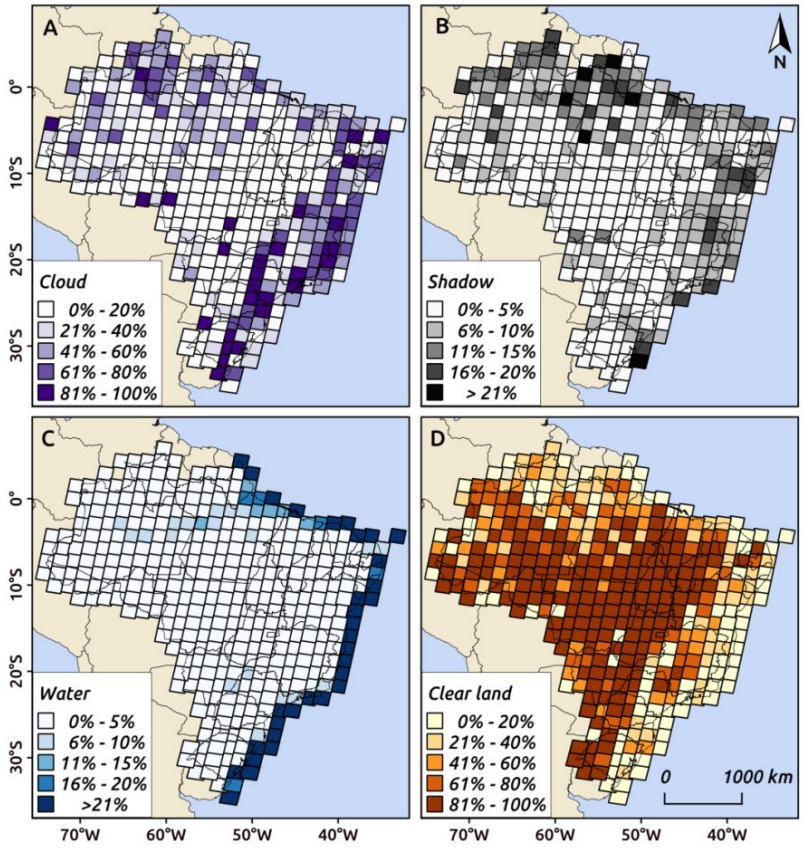
3. RESIIM Results and Discussion

3.1 ReSIIM results along Brazil

As expected, targets of *cloud*, *cloud shadow* and *water* were mainly found in coastal areas whilst in a contrary manner in those regions the scenes presented less clear land (Figure 5).

Considering that *forest* was based on Amazon forest (NDVI ≥ 0.8), this target was found mainly along Brazilian Amazon, indicating that for this kind of forest the threshold was suitable (Figure 6-A). Probably due to different types of native vegetation, *bare soil* was found mainly in Cerrado and Caatinga biomes in northeast Brazil (Figure 6-B).

Figure 5 – ReSIIM results for *cloud* (A), *shadow* (B), *water* (C) and *clear land* (D)

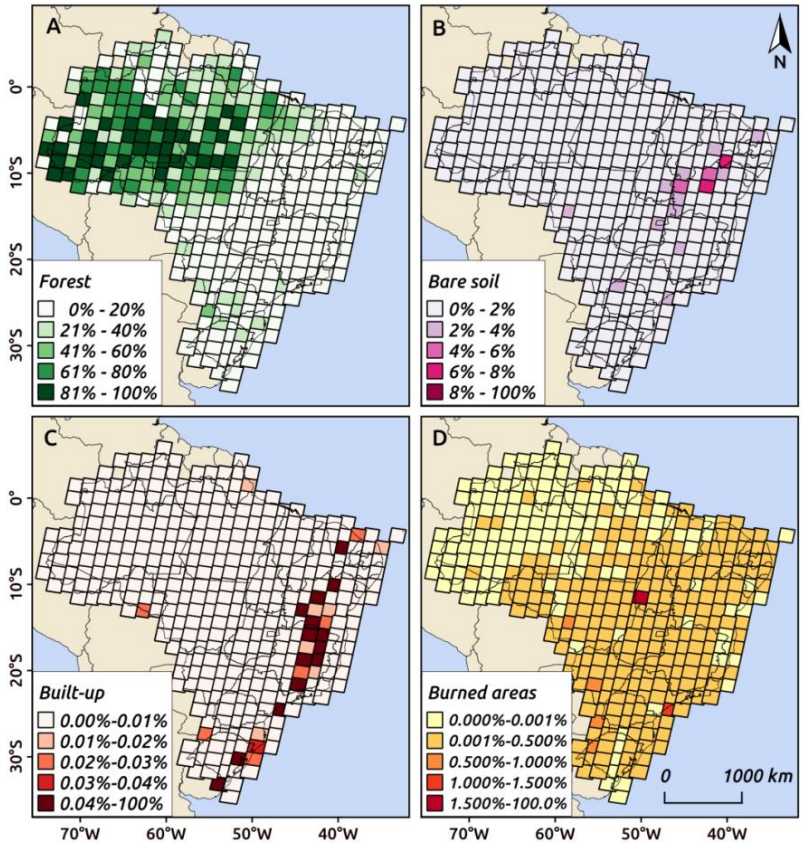


Source: Elaborated by the authors.

As a consequence of the rule used, none of the RS scenes represented more than 1% of the target *built-up*, which may indicate that part of these targets were probably filtered along with barren areas (Figure 6-C). Besides that, the remaining *built-up* areas also presented misclassification areas, probably with barren areas and due to the dry season, which may be indicated by the *built-up* detection along Minas Gerais and Bahia States. Finally, as expected, this target is also identified São Paulo State, more precisely São Paulo municipality, once NDBI accuracy is lowered in peripheral urban areas (ZHA et al., 2003).

For *burned* areas, we integrated two indices in order to restrict the results. In such a way, the most burned scene found by ReSIIM was 223/068, located in Tocantins state. According to *Queimadas Project* (INPE, 2018), this scene presents 80,401 ha of burned area, which represents about 2% of scene.

Figure 6 – ReSIIM results for *forest* (A), *bare soil* (B), *built-up* (C) and *burned areas* (D)



Source: Elaborated by the authors.

3.2 ReSIIM results for burned forest detection

The technique LSMM is the most accurate up to now for burn detection, however, as a drawback, it can be only applied in suitable images, which are free of clouds and with the potential to show burned areas. As this process is currently done manually, in this Section we verify ReSIIM's potential as a support tool for LSMM application. For that, some tests were performed aiming to retrieve from a database only appropriate images for LSMM application.

The tests were based only on the basic attributes available in Fmask algorithm, focusing on phenomena in lieu of date or location. Summing up 60 Landsat scenes, representing almost 200 million hectares and about 120 GB of data, two main datasets were selected to compose the database, 30 appropriate images (API) and 30 inappropriate images (INI). API refer to images with potential to present burned areas, while INI, do not. The aim of the tests is to comprehend which combination of metadata information is needed to maximize API and minimized INI retrieval.

Through the tests, we were able to comprehend better not only the available data, but also the role of ReSIIM in burned forest detection. Firstly we assessed individual targets in order to support image retrieval. After that, the targets with best performance were combined and evaluated as well. An overview of the main tests is present in Table 2.

Table 2 – Tests aiming to identify ReSIIM's potential as a support tool for LSMM application. Blue cells: equal or less than; Orange cells: equal or greater than.

Test number	Search Criteria	Percentage										Number of Retrieved Images	
		10	20	30	40	50	60	70	80	90	100	Appropriate (API)	Inappropriate (INI)
1	*cloud*											27	0
2	*cloud*											30	1
3	*shadow*											30	29
4	*clear land*											0	30
5	*clear land*											30	0
6	*clear land*											26	0
7	*water*											30	29
8	*cloud*											30	0
	clear land												

Source: Elaborated by the authors.

Although clouds are crucial targets for image retrieval in burned forest detection studies, tests 1 (percentage of *clouds* in scenes is not superior to 10%) and 2 (percentage of *clouds* is not superior to 20%) showed that it is not the only important searching criterion, once none of them could retrieve all API without the presence of INI. That probably happens because in some areas of the Brazilian Amazon, it is not possible to access RS images without the presence of *clouds* along the year, due to local humidity. *Shadow* targets were not relevant in the process, since almost 100% of the dataset is composed of images with less than 10% of *shadow* (test 3). In this manner, the range used in ReSIIM (pace of 10%) is not fine enough to enable the use of this target as a potential search criterion.

As well as *cloud*, in tests 4-6, it is possible to notice that clear land also plays an important role in image retrieval tests. Images with more than 30% of *clear land* (test 5) retrieved all the API and none of the INI. Although this is the best-case scenario, it is indicated to integrate the target *cloud* in the

process. Considering that *clear land* refers to the object of the study, once forests and burning processes occur on this kind of surface, the key issue for including *cloud* in the search criterion is the possibility to filter the results, as already highlighted by Pletsch et al. (2018).

Besides that, the target *water* in the test 7 was not able to discriminate both groups. It may have occurred once some areas present water along the whole year, thus this kind of filter would be more suitable if applied for more detailed and specific studies, such as along coastlines.

Finally, we combined *cloud* and *clear land* in the searching criteria (test 8), once those were the main targets identified in the aforementioned steps. The intersection between both was satisfactory, once all the images from API and none of the INI were retrieved. In such a way, the retrieved data were free of clouds and presented the potential to show burned areas.

4. Conclusions

The Remote Sensing data deluge is overwhelming the capacity of institutions to manage and retrieve its content. In this context, ReSIIM is a fast and easy alternative tool for Remote Sensing image information mining. It is based on the application of well-known methods for information extraction from RS scenes, storing this information and allowing users to access land use and land cover metadata through open source software, scripts and libraries. The developed tool will support many other avenues of sustainable research, considering that it enables phenomena searching criteria in lieu of just location or date parameters, as available in current official catalogs.

The use of spectral indices thresholds or the definition of phenomena such as forests are complex and require adaptations according to the location and application. For burned area detection in Brazilian Amazon, LSMM is yet required due to its high accuracy. In this context, ReSIIM is a support tool for its application, once the developed tool was able to retrieve all reference

data from the database. The crucial targets for our test application in burned forest detection were *cloud* and *clear land,* filtering noises and focusing on the object of study. More analyses are although required in order to combine different targets and Remote Sensing image retrieval results for further understanding of Earth phenomena, such as LULC changes.

ReSIIM methodology accuracy is not absolute, since it is an indication of target correlations using mathematical models. However, ReSIIM can be continuously improved according to its user's requirements. In this context, future research is also required in order to comprehend user's demands.

Acknowledgement

We would like to thank National Council for Scientific and Technological Development (CNPq) for the research financial support, processes 131241/2016-8 and 140377/2018-2, and the grant #2017/24086-2, from São Paulo Research Foundation (FAPESP).

References

ANDERE, L.; ANDERSON, L. O.; DUARTE, V.; ARAI, E.; ARAGÃO, J. R. L.; ARAGÃO, L. E. O. C. Dados multitemporais do sensor MODIS para o mapeamento de queimadas na Amazônia. **Proceedings of the XVII Simpósio Brasileiro de Sensoriamento Remoto (SBSR)**, João Pessoa, 2015. pp. 3534-3541.

ANDERSON, L. O., SHIMABUKURO, Y. E., DEFRIES, R. S., & MORTON, D. Assessment of deforestation in near real time over the Brazilian Amazon using multitemporal fraction images derived from Terra MODIS. **IEEE Geoscience and Remote Sensing Letters**, v. 2, n. 3, 2005a. pp. 315-318.

ANDERSON, L. O.; ARAGÃO, L. E. O. C.; LIMA, A. D.; SHIMABUKURO, Y. E. Detecção de cicatrizes de áreas queimadas baseada no modelo linear de mistura espectral e imagens índice de vegetação utilizando dados multitemporais do

sensor MODIS. TERRA no Estado do Mato Grosso, Amazônia Brasileira. **Acta Amazonica**, v. 35, n. 4, 2005b. pp. 445-456.

ANDERSON, L. O.; ARAGÃO, L. E. O. C.; GLOOR, M.; ARAI, E.; ADAMI, M.; SAATCHI, S. S.; MALHI, Y.; SHIMABUKURO, Y. E.; BARLOW, J.; BERENGUER, E.; DUARTE, V. Disentangling the contribution of multiple land covers to fire-mediated carbon emissions in Amazonia during the 2010 drought. **Global biogeochemical cycles**, v. 29, n. 10, 2015. pp. 1739-1753.

ARAGÃO, L. E. O.; MALHI, Y.; ROMAN-CUESTA, R. M.; SAATCHI, S.; ANDERSON, L. O.; SHIMABUKURO, Y. E. Spatial patterns and fire response of recent Amazonian droughts. **Geophysical Research Letters**, v. 34, n. 7, 2007.

ARAGÃO, L. E. O. C.; ANDERSON, L. O.; LIMA, A.; ARAI, E. Fires in Amazonia. In: **Interactions Between Biosphere, Atmosphere and Human Land Use in the Amazon Basin**. Springer Berlin Heidelberg, 2016. pp. 301-329.

ARAGÃO, L. E. O. C.; SHIMABUKURO, Y. E. The incidence of fire in Amazonian forests with implications for REDD. **Science**, v. 328, n. 5983, 2010. pp. 1275-1278.

BANNARI, A.; MORIN, D.; BONN, F.; HUETE, A. R. A review of vegetation indices. **Remote sensing reviews**, v. 13, n. 1-2, 1995. pp. 95-120.

BASTARRIKA, A.; CHUVIECO, E.; MARTÍN, M. P. Mapping burned areas from Landsat TM/ETM+ data with a two-phase algorithm: Balancing omission and commission errors. **Remote Sensing of Environment**, v. 115, n. 4, 2011. pp. 1003-1012.

CARDOZO, F. S.; PEREIRA, G.; SHIMABUKURO, Y. E.; MORAES, E. C. Análise do uso do Modelo Linear de Mistura Espectral (MLME) para o mapeamento das áreas queimadas no Estado de Rondônia no ano de 2010. **Proceedings of the XVI Simpósio Brasileiro de Sensoriamento Remoto (SBSR)**, Foz do Iguaçu, 2013. pp. 7265-7272.

CHI, M.; PLAZA, A.; BENEDIKTSSON, J. A.; SUN, Z.; SHEN, J.; ZHU, Y. Big Data for Remote Sensing: Challenges and Opportunities. **Proceedings of the IEEE**, v. 104, n. 11, 2015. pp. 2207–2219.

CHUVIECO, E.; CONGALTON, R. G. Mapping and inventory of forest fires from digital processing of TM data. **Geocarto International**, v. 3, n. 4, 1988. pp. 41-53.

DATCU, M.; SEIDEL, K.; PELIZARRI, A.; SCHROEDER, M.; REHRAUER, H.; PALUBINSKAS, G.; WALESSA, M. Image information mining and remote sensing data interpretation. **Proceedings of the IGARSS 2000 - IEEE 2000 International Geoscience and Remote Sensing Symposium**. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120), v. 7, n. July, 2000. pp. 3057–3059.

DATCU, M.; SEIDEL, K. An Innovative Concept for Image Information Mining. **Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining**. Springer, Berlin, Heidelberg, 2002. pp. 84-99.

DATCU, M.; SEIDEL, K. Image information mining : exploration of Earth observation archives. **Geographica Helvetica**, v. 58, n. 2, 2003. pp. 154–168.

EAKINS, J. P.; GRAHAM, M. E. **Content based image retrieval**. A report to the JISC technology applications programme, Institute for Image Data Research, University of Northumbria, Newcastle, 1999.

FAO, FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS; ITTO, INTERNATIONAL TROPICAL TIMBER ORGANIZATION. **The State of Forests in the Amazon Basin, Congo Basin and Southeast Asia**. A report prepared for the Summit of the Three Rainforest Basins Brazzaville, Republic of Congo, 80 p., 2011.

FLOOD, N.; GILLINGHAM, S. **Python Fmask Documentation - Release 0.4.5**. Available in: <https://media.readthedocs.org/pdf/pythonfmask/latest/pythonfmask.pdf>. Accessed in July 2018.

FRANK, S. Cataloging digital geographic data in the information infrastructure: A literature and technology review. **Information Processing and Management**, v. 30, n. 5, 1994. pp. 587–606.

GAO, X., HUETE, A. R., NI, W.; MIURA, T. Optical–biophysical relationships of vegetation spectra without background contamination. **Remote Sensing of Environment**, v. 74, n. 3, p. 609-620, 2000.

GARCÍA, M. J. L.; CASELLES, V. Mapping burns and natural reforestation using thematic Mapper data. **Geocarto International**, v. 6, n. 1, 1991. pp. 31–37.

GOODRUM, A. A. Image information retrieval: An overview of current research. **Informing Science**, v. 3, n. 2, 2000. pp. 63–66.

GUPTILL, S. C. Metadata and data catalogues. **Geographical Information Systems: Principles and Technical Issues**. John Wiley & Sons, v. 2, 1999. pp. 677–692.

HUETE, A., DIDAN, K., MIURA, T., RODRIGUEZ, E. P., GAO, X.; FERREIRA, L. G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n. 1-2, p. 195-213, 2002.

INPE, INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. **Projeto PRODES**: Monitoramento da floresta amazônica brasileira por satélite. Available in: <http://www.obt.inpe.br/prodes/index.php>. Accessed in June 2017.

INPE, INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. **Projeto Queimadas: Área Queimada 30m**. Available in: <http://prodwww-queimadas.dgi.inpe.br/aq30m/>. Accessed in June 2018.

JIN, X. M.; ZHANG, Y. K.; TANG, Y.; HU, G. C.; GUO, R. H. Quantifying bare soil evaporation and its relationship with groundwater depth. **International Journal of Remote Sensing**, v. 35, n. 21, 2014. pp. 7567–7582.

KEY, C.; BENSON, N. **General technical report: Landscape assessment (LA) - sampling and analysis methods**. United States: US Department of Agriculture, USDA. Forest Service, 2006. RMRS-GTR-164-CD.

KÖRTING, T. S.; NAMIKAWA, L. M.; FONSECA, L. M. G.; FELGUEIRAS, C. A. How to effectively obtain metadata from remote sensing big data?. **Proceedings of GEOBIA 2016**. SOLUTIONS AND SYNERGIES, Enschede, Nederlands. Available in : <http://https://www.conftool.net/geobia2016/index.php?page=browseSessions\&abstracts=show\&form_session=16\&presentations=show>. Accessed in March 2018.

LANEY, D. 3D data management: Controlling data volume, velocity and variety. Controlling data volume, velocity and variety. **META group research note**, v. 6, n. 70, 2001. pp. 1.

LASHOF, D. A. The contribution of biomass burning to global warming: An integrated assessment. **Proceedings of Global biomass burning: atmospheric, climatic, and biospheric implications**. Williamsburg, VA (United States): Massachusetts Inst. of Tech. Press, 1991. Available in:

<https://inis.iaea.org/search/search.aspx?orig_q=RN:23067069>. Accessed in July 2018

LI, J.; NARAYANAN, R. M. Integrated Information Mining and Image Retrieval in Remote Sensing. In: CHANG, C. I. (Ed.). **Recent Advances in Hyperspectral Signal and Image Processing**. 1. ed. Trivandrum, India: Transworld Research Network, 2006. pp. 449--478.

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A.; CHENG, T. Geospatial big data handling theory and methods: A review and research challenges. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 115, 2016. pp. 119–133. ISSN 09242716.

LIANG, S.; ZHANG, X.; XIAO, Z.; CHENG, J.; LIU, Q.; ZHAO, X. **Global LAnd Surface Satellite (GLASS) products: algorithms, validation and analysis**. Berlin, Germany: Springer Science & Business Media, 2013.

LONG, F.; ZHANG, H.; FENG, D. D. Fundamentals of Content-Based Image Retrieval. In: FENG, D. D.; SIU, W.-C.; ZHANG, H.-J. (Eds.). **Multimedia Information Retrieval and Management: Technological Fundamentals and Applications**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 1–26.

MARTÍN ISABEL, M. D. P. Cartografía de grandes incendios forestales en la Península Ibérica a partir de imágenes NOAA-AVHRR. **Série geográfica**, n. 7, 1998. pp. 109-128. ISSN: 1136-5277.

MATSUSHITA, B., YANG, W., CHEN, J., ONDA, Y.; QIU, G. Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: a case study in high-density cypress forest. **Sensors**, v. 7, n. 11, p. 2636-2651, 2007.

NAMIKAWA, L. M.; CASTEJON, E. F. **Mapas de Lamina de Água para Todo o Brasil Extraídos do RapidEye**. Available in: <http://wiki.dpi.inpe.br/doku.php?id=mapas:waterbodies>. Accessed in June 2017.

NAMIKAWA, L. M.; KÖRTING, T. S.; CASTEJON, E. F. Water Body Extraction from RapidEye Images: An Automated methodology based on hue Component

of Color Transformation from RGB to HSV Model. **Revista Brasileira de Cartografia**, v. 68, n. 6, 2016. pp. 1097–1111.

NOAA, NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. A. **MGVI - Normalized Difference Vegetation Index**. Available in: <http://www.ospo.noaa.gov/Products/land/mgvi/NDVI.html>. Accessed in June 2017.

PADILLA, M.; OLOFSSON, P.; STEHMAN, S. V.; TANSEY, K. Stratification and sample allocation for reference burned area data. **Remote Sensing of Environment**, Elsevier, v. 203, 2017. pp. 240–255. ISSN 0034-4257.

PLETSCH, M. A. J. S.; KÖRTING, T. S. Remote Sensing Image Information Mining applied to Burnt Forest Detection in the Brazilian Amazon. **Proceedings of XVIII GEOINFO**, Salvador, BA, Brazil, 2017. Available in: <http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m16c/2017/12.01.20.06/doc/40pletsch_korting.pdf>. Accessed in June 2018.

PLETSCH, M. A. J. S.; PENHA, T. V.; JUNIOR, C. H. L. S.; KÖRTING, T. S.; ARAGÃO, L. E. O. C.; ANDERSON, L. O. Integração do algoritmo FMASK ao modelo linear de mistura espectral como subsídio à detecção de áreas queimadas na Amazônia brasileira. **Revista Brasileira de Cartografia**, v. 70, n. 2, 2018. pp. 696-724.

QUARTULLI, M.; OLAIZOLA, I. G. A review of EO image information mining. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 75, 2013. pp. 11–28. ISSN 0924-2716.

RIKIMARU, A.; ROY, P. S.; MIYATAKE, S. Tropical forest cover density mapping. **Tropical Ecology**, v. 43, n. 1, 2002. pp. 39–47.

ROUSE, J. W. Jr.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. **Proceedings of Earth Resources Technology Satellite-1 Symposium**, Washington, D.C.: NASA. Goddart Space Flight Center, v.1, 1973. pp. 309-317. (NASA SP-351). Available in: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19740022614.pdf>. Accessed in December 2018.

SHIMABUKURO, Y. E.; DUARTE, V.; ARAI, E.; FREITAS, R.; LIMA, A.; VALERIANO, D.; BROWN, I.; MALDONADO, M. Fraction images derived from terra modis data for mapping burnt areas in Brazilian Amazonia.

International Journal of Remote Sensing, Taylor & Francis, v. 30, n. 6, 2009. pp. 1537–1546.

SHIMABUKURO, Y. E.; MIETTINEN, J.; BEUCHLE, R.; GRECCHI, R. C.; SIMONETTI, D.; ACHARD, F. Estimating burned area in Mato Grosso, Brazil, using an object-based classification method on a systematic sample of medium resolution satellite images. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 8, n. 9, 2015. pp. 4502–4508. ISSN 21511535.

SHIMABUKURO, Y. E.; SMITH, J. A. The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 29, n. 1, 1991. pp. 16–20. ISSN 15580644.

SHYU, C. R. et al. GeoIRIS: Geospatial Information Retrieval and Indexing System mdash;Content Mining, Semantics Modeling, and Complex Queries. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, n. 4, 2007. pp. 839–852.

STEPINSKI, T. F.; NETZEL, P.; JASIEWICZ, J. LandEx - A geoweb tool for query and retrieval of spatial patterns in land cover datasets. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 7, n. 1, 2014. pp. 257–266. ISSN 19391404.

ZHA, Y.; GAO, J.; NI, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. **International Journal of Remote Sensing**, v. 24, n. 3, 2003. pp. 583–594.

ZHU, Z.; WANG, S.; WOODCOCK, C. E. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4,7, 8, and Sentinel 2 images. **Remote Sensing of Environment**, v. 159, 2015. pp. 269–277.

ZHU, Z.; WOODCOCK, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. **Remote Sensing of Environment**, v. 118, 2012. pp. 83–94.