

## Mass movements' scars classification using data mining techniques

Jéssica Gerente<sup>1</sup>  
Camile Söthe<sup>1</sup>  
Priscila Negrão<sup>1</sup>  
Thales Sehn Körting<sup>1</sup>

<sup>1</sup>Instituto Nacional de Pesquisas Espaciais - INPE  
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil  
{ jessica.gerente, camile.sothe, priscila.negrao, thales.korting }@inpe.br

**Abstract.** Mass movements are destructive natural phenomena that can lead to serious problems such as economic loss, damage to natural resources and even injuries and deaths. Efforts have been made to semi automate the interpretation of remote sensing data in order to improve efficiency and support specialists in recognizing mass movements' scars. However, this approach is still incipient in Brazil. This study presents results of semiautomatic classification of mass movements' scars that occurred in Nova Friburgo (Rio de Janeiro state, Brazil) in 2011 by using segmentation and applying data mining techniques. Two classifications were compared, from C4.5 and CART decision tree algorithms. Data mining techniques confirmed that mass movements have different spectral characteristics from other classes, allowing its detection from remote sensing images. The overall accuracy of C4.5 algorithm was 62.6%, while CART was 66.4%. The errors occurred mainly in urban areas and in unpaved roads located at higher altitudes. Spectral digital elevation model (DEM) average, blue band and NDVI were the more appropriate attributes to distinguish mass movements patterns. This methodology offered an alternative, that still needs improvements, to produce data about statistics and spatial distribution of mass movements, providing information to be used, for instance, as parameters in susceptibility maps and models, assisting public policies focused on natural disasters.

**Keywords:** decision tree, features, natural disasters

### 1. Introduction

Mass movements are destructive natural phenomena that can lead to serious problems such as economic loss, damage to natural resources and even injuries and deaths (Klose et al., 2014). In Brazil, there has been a significant increase in the frequency of natural disasters in the past few decades particularly associated with mass movements (UFSC, 2013). For some authors this increase is mainly due to climate change (increase in extreme events) and the growth of irregular settlements in urban areas (Robaina, 2008; Coelho Netto, 2011).

Studies on mass movements are important because they provide an analysis of the associated risks. Usually the parameters for mass movements' mapping are derived from historical data, field surveys and visual interpretation of satellite or aerial images. However, historical data are not always available or complete, intensive field research is impractical for studies on large scales, and visual analysis of spectral images can be a time consuming task (Escape et al., 2014), susceptible to analyst's subjectivity.

In this context, the attribute extraction and classification of regions where typical mass movements' signals are present in orbital images is a challenging task. The difficulty associated with this procedure is mainly due to the great complexity of shape, size, texture, and other variables related to this hillslope process (Selby, 1993). Thus, efforts have been made to semi automatize the interpretation of remote sensing data in order to improve efficiency and support specialists in recognizing mass movements' scars (Martha et al., 2012). However, this approach applied to mass movements' scar detection is still incipient in Brazil.

Data mining includes a set of techniques to extract useful information from a database with a large volume of data through intelligent methods. According to Körting et al. (2012), data mining techniques can increase the potential for analysis and applications of remote sensing data, once they present a great diversity of targets that are difficult to distinguish and, therefore,

require better techniques for extracting information. Hence, data mining allows the classification of images in a quicker way, compared to manual analysis.

The model derived from data mining can be represented in several forms, such as decision trees, which consists on a tree structure which can be converted into classification rules (Han and Kamber, 2006). They have been widely used because of their intuitive representation, which makes the classification model easier to be interpreted. According to Quinlan (1993), the classification by decision trees has the advantage of owning non-parametric properties and is capable of classifying images with statistic distributions different from the Gaussian, including heterogeneous and noisier data, as outliers.

Therefore, the aim of this study is to classify semi automatically scars of mass movements that occurred in the natural disaster of January 2011 in Nova Friburgo city (Rio de Janeiro state, Brazil), using image segmentation and data mining techniques, identifying the best attributes that differentiate these areas from other types of land use and land cover classes.

## 2. Material and Methods

The study area has 41.25 km<sup>2</sup> and is located in the Roncador River basin, at the district of Córrego Dantas, Nova Friburgo-RJ, Brazil. This county is located in Rio de Janeiro's mountain region, a geomorphological unit known as "*Planalto Reverso da Região Serrana*" (Dantas, 2001), with steep mountainous terrain and altitude ranging from 400 to 2.300m. The steeper and higher terrains are covered with primary/preserved forests, totaling about 70% of its territory (CIDE, 2003). Due to such terrain settings, this geomorphological unit has a high susceptibility to erosion events such as mass movements (Dantas, 2001).

As methodological procedures, it was used topography data from the TOPODATA project (Brasil, 2008), with 30m of spatial resolution. Also it was used the scene n° 2328825 from the RapidEye constellation sensor acquired in August 13<sup>th</sup>, 2011. This sensor has 5m of spatial resolution and 5 spectral bands, three of them in the visible area, one in near infrared (NIR) and one situated on the edge of the red (Red-edge). Due to the relative short period of time between the disaster of January 2011 in the *Serrana* region of Rio de Janeiro and the date of acquisition of the images, it was possible to classify the scars of mass movements, which were still apparent on the terrain. In addition, the Normalized Difference Vegetation Index (NDVI) was calculated to assist in the segmentation process, since the index contrasts mass wasting areas with other targets, especially vegetation.

For validation purposes, the scars of mass movements appearing in the study area were vectorized in the software QGIS 2.8 by visual interpretation. As an ancillary data for recognizing of mass movements' occurrences, historical high spatial resolution images from Google Earth software were analyzed.

For data preprocessing, it was performed atmospheric correction of the RapidEye scene, using Quick Atmospheric Correction algorithm (QUAC) available at ENVI 5.0. Still in ENVI 5.0, the following procedures were executed: 1) resizing the original image to an area of interest containing 1422 x 1167 pixels; 2) NDVI index calculation; and 3) layer stack of RapidEye multispectral bands (1 to 5) with DEM (6) and NDVI (7), respectively.

The process of classification was first performed in the software InterIMAGE 1.43 (Costa et al., 2008). Initially, a semantic network was built representing the classes expected to be found in the scene. In this paper, operational networks were created with no hierarchical relationship between classes, since the objective was to explore the semi-automatic classification with C4.5. Thus, each class (leaf node) was associated with the same parent node, without intermediate levels. As result, five classes were stipulated: *mass movements*, *urban area*; *vegetation*; *field* (comprising *grass*, *agriculture* and *meadow* areas) and *rock*. Using the Samples Editor tool, segmentation was performed in the images using the multiresolution Region Growing algorithm, proposed by Baatz and Schäpe (2000). This algorithm provides four

user-defined parameters: color, shape, scale factor and weight for each band used in the segmentation. Following the same methodology applied by Francisco and Almeida (2013) for visual choice of scale parameters, three consecutive segmentation levels were performed, reducing the scale factor in the performance of each new proceeding, creating a new level with the largest number of small-sized objects. Also, the weight given to the spectral bands and the NDVI layer were varied until an appropriate result was found to delimit the mass movement scars.

After the segmentation, samples were randomly collected. We collected 100 samples of the mass movement class and 159 samples distributed among the other established classes. Then, attribute extraction was performed for each segment to be used in the classification. A total of 51 attributes were created, from which, 43 were spectral parameters obtained for each spectral band used (except brightness that was obtained for the entire scene); 04 were operations of spectral bands, including NDVI and Simple Ratio Index (SR), and the other 04 were spatial attributes (Table 1). More information about the attributes generated can be obtained by Silva et al. (2005) and Körting (2012).

**Table 1.** Extracted attributes of each segment in InterIMAGE.

Attribute	Type
Mean (bands 1-6)	Spectral
Entropy (bands 1-6)	Spectral
Band Ratio (bands 1-6)	Spectral
Standard Deviation (bands 1-6)	Spectral
Amplitude (bands 1-6)	Spectral
Minimum pixel value (bands 1-6)	Spectral
Maximum pixel value (bands 1-6)	Spectral
Brightness	Spectral
NDVI = (NIR - Red) / (NIR+ Red)	Operation
(NIR - Red)	Operation
Simple Ratio = (NIR/ Red)	Operation
(NIR/Red-edge)	Operation
Shape Index	Spatial
Bounding Box área	Spatial
Perimeter Area Ratio	Spatial
Compacity	Spatial

The classification was made with TA\_C45\_Classifier top down algorithm, which uses the concept of the decision tree proposed by Quinlan (1993). The final classification results were generated in a vector data in the format *shapefile* (.shp), while the decision tree is generated in a *text* (.txt) file.

In order to explore the attributes that better distinguish the mass movements' class of other targets, it was still used the data mining tool WEKA 3.7, which incorporates a set of machine learning algorithms to enable the extraction of knowledge. The methodology developed in WEKA comprise the view of attributes generated in InterIMAGE through scatterplots, and a classification model generation for the CART decision tree (Classification and Regression Trees), proposed by Breiman et al. (1984). The database used in this work was the samples and their attributes generated in InterIMAGE and converted to the Attribute- Relation File Format (.ARFF).

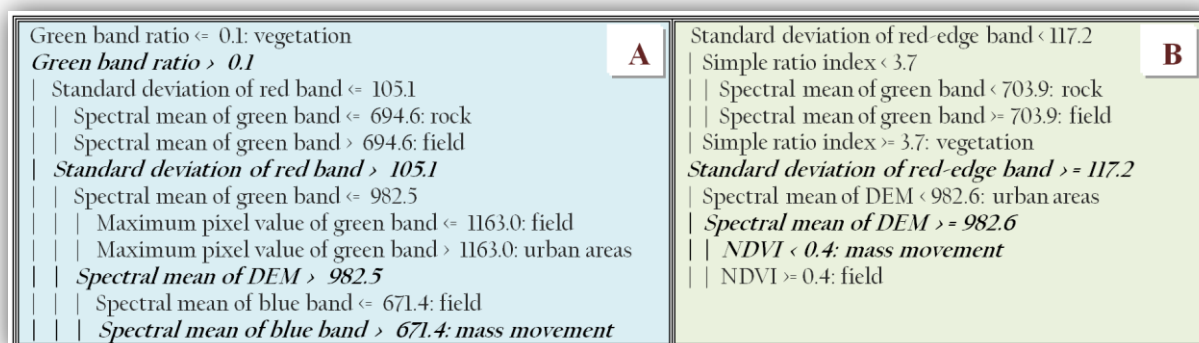
To evaluate the accuracy, the classified maps were crossed with the reference map, and from that, it was obtained the number of areas considered as true positive (TP), false negative (FN) and false positive (FP). TPs comprise the correctly mapped mass movements, whereas the other two identification categories represent two types of identification errors. FNs correspond to reference mass movements that have not been identified by the approach, and FPs are identified as mass movements objects which, again, have not been mapped in the reference inventory (Martha et al., 2012). Based on this relation, it was used three accuracy metrics:

*Detection Percentage, Omission Error, Commission Error.* *Detection Percentage* represents the percentage of mass movements which have been correctly identified by the automated approach. *Commission Error* and *Omission Error* describe separately the influence of the two possible identification errors FP and FN, respectively (Behling et al., 2014).

### 3. Results and Discussion

In segmentation, the most suitable scale parameter for visually delimiting mass movements areas was 60, with weight 0.5 assigned to the red, NIR bands and NDVI. For other bands, it was attributed the weight 0. It is noteworthy that, by providing a greater weight to these bands or including more bands in the segmentation process, the tendency of the algorithm is to over-segment the image. Color and shape parameters, as mentioned above, were kept at the value of 0.5. These parameters, although generating a larger number of segments for the same mass movement scar, were used to prevent that classes with similar spectral response (such as some field and urban areas) were included in the same segment as mass movement class.

Figure 1 presents the decision tree obtained with each algorithm, C4.5 from InterIMAGE and SimpleCart from WEKA.



**Figure 1.** Decision tree generated by the algorithm C4.5 (A) and by the algorithm CART (B). Highlighted, the attributes and rules used to classify mass movements of other classes.

In the decision tree generated by the C4.5, the root node was the ratio of the band corresponding to the spectral mean of green band. This attribute was used to first separate vegetation class of other classes. The second attribute was the standard deviation of the red spectral band. This attribute discriminated classes with lower standard deviation, rock and field, from classes with higher standard deviations, such as mass movement, urban areas and field. Field class appeared in both branches, for having certain spectral range characterized by areas being covered with sparse vegetation and areas where the soil was more exposed.

Urban areas and mass movements are also spectrally similar classes with high reflectance in the visible and NIR. To differentiate these three classes, the algorithm selected the spectral mean of DEM band. Lower values of spectral mean of DEM band discriminated field and urban classes from field and mass movements' classes. This choice is justified by the fact that mass movements are situated in higher notable places than urban areas, for instance. However, the field class is located both in high and lower areas, appearing in both branches.

Finally, in the last branch, remaining samples of the field class were differentiated of mass movements with the spectral mean blue band. The presence of herbaceous vegetation in certain field areas reduces its spectral response at wavelengths corresponding to blue, because of photosynthetic pigments in vegetation. On the other hand, bare soil, which are the areas of mass movements, has a comparatively higher response than the vegetation in this spectrum band. Thus, lower values of spectral mean of blue band were diagnosed as field while larger values were classified as mass movements.



For the tree generated by CART algorithm, the root node was the Red-Edge band standard deviation, whereas lower values discriminated the vegetation, field and rock classes, and higher values designated the urban, field and mass movement classes. As in the C4.5, the CART algorithm also selected the spectral mean of DEM attribute to discriminate the urban class of mass movement and field classes, assigning lower values of this attribute to the urban class. Unlike other algorithm, to differentiate field of mass movements, CART selected NDVI, considering that higher values are equivalent to field class and lower values are equivalent to mass movement class. Such behavior is expected, given that the vegetation under the field class gives higher values of NDVI for this class.

Finally, areas classified as mass movements were extracted into a shapefile to be compared with the reference. The decision tree C4.5 resulted in a 285 ha area classified as mass movement while the CART algorithm resulted in an area of 311 ha. It is noticed that both classifications overrated the reference that had 260 ha of mass movement areas. Nevertheless, in strictly numerical terms, that is, aiming only the quantification of mass movement area and not its localization, it can be assumed that the C4.5 algorithm was reasonable, overestimating ~15 ha in relation to the reference. Table 2 shows the results obtained with cross tabulation and accuracy evaluation. Similar results were found by Hölbling et al. (2015) that also had a classification with overestimated results in relation to the reference map. The authors classified mass movements resulting from two typhoons using the object-based approach in SPOT-5 images in a 16 km<sup>2</sup> area in north Taiwan.

**Table 2.** Evaluation of C4.5 and CART algorithms in relation to the reference.

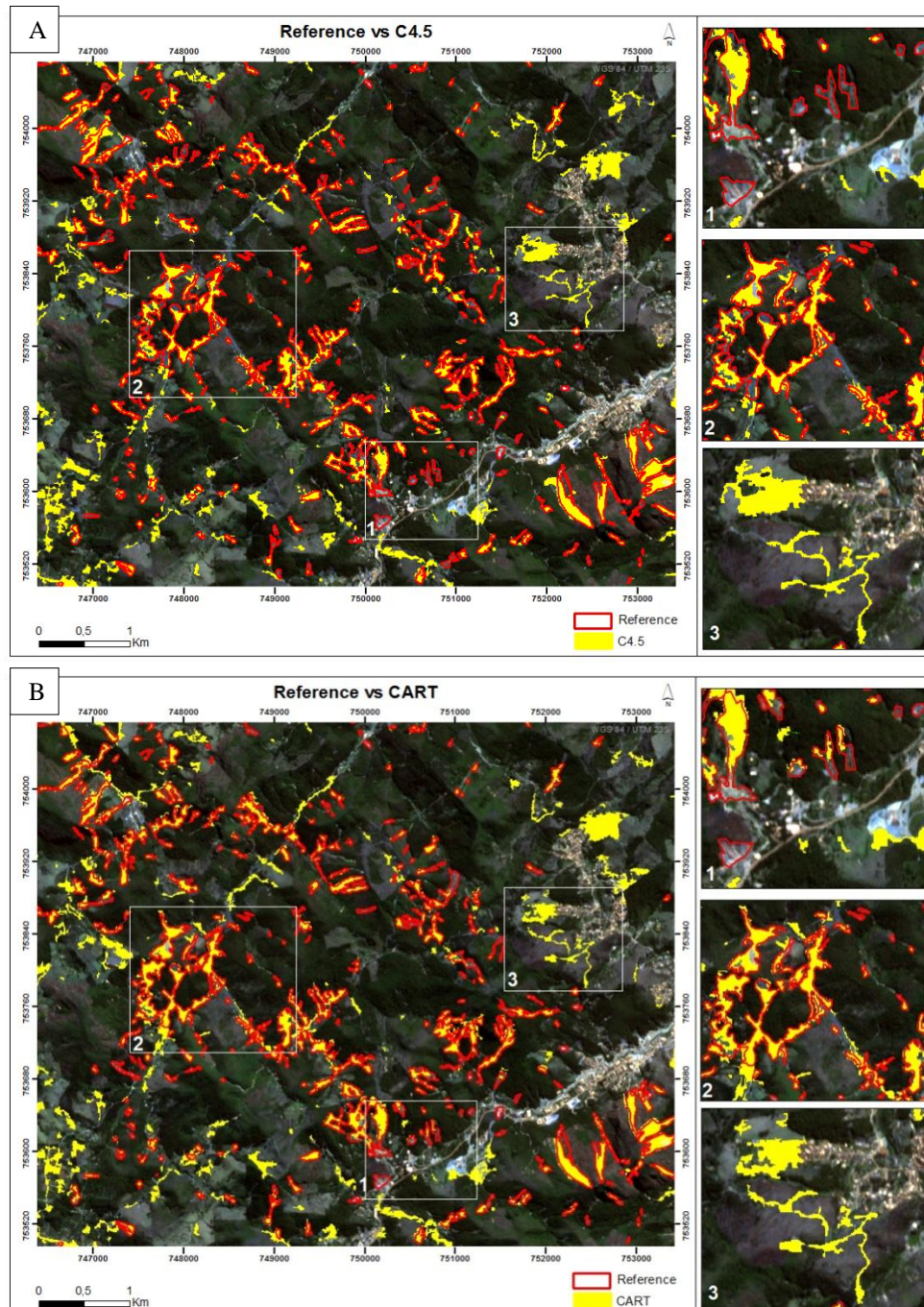
Accuracy Assessment	Algorithm	
	C4.5	CART
True positive (ha)	164	174
False negative (ha)	98	88
False positive (ha)	121	137
Total classification area (ha)	285	311
<b>Detection percentage (%)</b>	<b>62.6</b>	<b>66.4</b>
<b>Omission error (%)</b>	<b>37.4</b>	<b>33.6</b>
<b>Commission error (%)</b>	<b>42.5</b>	<b>44.1</b>

It is seen in Table 2 that both algorithms had a detection percentage of mass movement scars areas superior than 60%. Although the C4.5 algorithm has generated a total area value more similar to the reference, this algorithm had lesser detection percentage of mass movements (62,4%). In other words, comparing to CART algorithm, C4.5 generated a classification with larger spatial divergence in relation to the reference. The C4.5 algorithm had 37,4% of omission error and 42,5% of commission error. On the other hand, CART algorithm generated a detection percentage of 66,4%, an omission error of 33,6% and a commission error of 44,1%. Based in these values it can be inferred that CART algorithm is closer to the reference in relation to the areas classified as mass movements.

Both algorithms generated more commission errors than the omission ones due to the overestimation of the area classified as mass movement. The algorithms inaccurately classified some urban areas and unpaved roads as mass movements. In general, unpaved roads on slope areas are constructed in valley floors, once these areas have already been deforested and compacted in most cases. Valley floors are also regions where sediments flows often occurs; consequently, these areas are likely to form a “path” for mass movements. Therefore, both the spectral response and the form and altitude of these roads resemble mass movement’s scars, which may justify the algorithm errors in these areas.

In Figure 2 “A” and “B”, it can be noticed that both algorithms generated similar classifications. The main confusions were, as commented above, in urban areas, unpaved roads

and on valley floors. In the first zoom (1), it can be seen some mass movements scars with omission error, in other words, mass movement scars that were not detected by both algorithms. In addition, it is observed that the contrast between scars and vegetation is subtle. In the zoom frame number two; there are some areas with a good degree of accuracy in the classification. In the third zoom frame, one can see areas with commission errors which correspond to urban areas and to unpaved roads that were mistaken for mass movements.



**Figure 2.** Result of the classification of mass movements with the decision tree C4.5 (A) and CART (B) on the composition of R3G2B1 RapidEye image.

In Figure 2 it can be seen that both algorithms generated similar results in detecting mass movement scars. This similarity may be related to the fact that both algorithms used the attribute of spectral DEM mean to differentiate urban and mass movements' areas. A

hypothesis to explain the commission errors in urban areas is that some of these areas are located in higher altitudes (Figure 2, zoom 3). Also it can be noticed that some omission errors occurred at lower altitudes (Figure 2, zoom 1), that is, in places near the road and urbanized areas.

Still, in the studied image area, there are anthropic features such as cut slopes. These features are likely to generate confusion by the algorithm, once they have similar characteristics with mass movement scars: bare soil in higher altitudes (Figure 2, zoom 3). Another point to mention is that there was an overestimation by the algorithms in fluvio-colluvial plains and in riverbeds. Such environments are regions of transport and deposition of mass movement sediments. Due the complexity intrinsic to the mass movement process, as previously mentioned, mapping these features is a complex task especially because of the uncertainty related with what is in fact sediment or deposit of mass movements and what could be materials previously existent in the local.

Lastly, it is worth mentioning possible errors present in the reference map, once a field trip to collect data was not attempted. Some possible sources of errors in the reference map are: 1) scale problems during the mapping process, causing errors in the delimitation of the mass movements' scars borders; 2) possible errors of omission or commission derived from the visual interpretation process, which is subjective to the understanding of the interpreter; 3) the interpreter knowledge in relation to the surrounding features (that might help in the decision) combined with a time series of images available at Google Earth taken closer to the 2011 disaster and with a better spatial resolution. All these additional data and the own interpreter's intelligence are not available as algorithms attributes, increasing the possibilities of classification errors by the algorithms.

#### **4. Conclusions**

Data mining techniques pointed that mass movements have different spectral and spatial characteristics from other classes, allowing its detection from an orbital image. Although several attributes were computed from the samples, the decision tree algorithms only used six of them to make the classification of mass movement areas. As a result, data mining assists the analyst on the choice of attributes to differentiate classes. The manual analysis of each attribute would take time and could preclude some applications. Furthermore, the manual construction of decision thresholds would be subjected to the analyst and then it could be more difficult to replicate these methods in other study areas.

In this application, the attribute of spectral DEM average expressively contributed to mass movement differentiation from the urban area class, once both classes are spectrally similar in the visible and NIR. In addition, the spectral average of the blue band and NDVI helped the differentiation between mass movements and the field class, that also have some spectral similarity. Both algorithms had more than 60% accuracy. Larger commission errors were associated with urban area and unpaved roads classes situated at higher altitudes.

The decision tree offers the advantage of allowing the user to visualize the classification process. Also, its improvement could generate maps with better accuracy according to each area. Therefore, this methodology offered an alternative, that still can be improved, to produce data about statistics and spatial distributions of mass movements. It may provide information to be used, for instance, as parameters in susceptibility maps and models, assisting the management of public policies focused on natural disasters.

#### **Acknowledgment**

The authors are grateful to “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq) and “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) for the scholarships.



## References

- Baatz, M., Schäpe, M. **Multiresolution segmentation**—An optimization approach for high quality multi-scale image segmentation. In: Strobl, J., Blaschke, T., Griesebner, G. (Eds.), *Angewandte Geographische Informations-Verarbeitung XII*. Wichmann Verlag, Karlsruhe, p. 12–23, 2000.
- Behling, R.; Roessner, S.; Kaufmann, H.; Kleinschmit, B. Automated Spatiotemporal Landslide Mapping over Large Areas Using RapidEye Time Series Data. **Remote Sens.**, v. 6: 8026-8055, 2014.
- Brasil. Instituto Nacional de Pesquisas Espaciais (INPE). **Topodata: banco de dados geomorfométricos do Brasil. Variáveis geomorfométricas locais**. São José dos Campos, 2008. <<http://www.dsr.inpe.br/topodata/>>
- Breiman, L. et al. **Classification and Regression Trees**. Belmont, CA: Wadsworth. 1984.
- CIDE. Fundação Centro de Informações e Dados do Rio de Janeiro. **Índice de Qualidade de Municípios Verde II**. Rio de Janeiro: Secretaria de Estado de Planejamento, Desenvolvimento Econômico e Turismo, 2003. 154p.
- Costa, G. A. O. P., Pinho, C. M. D.; Feitosa, R. Q.; Almeida, C. M. de; Kux, H. J. H.; Fonseca, L. M. G.; Oliveira, D. A. B. INTERIMAGE: uma plataforma cognitiva open source para a interpretação automática de imagens digitais. **RBC. Revista Brasileira de Cartografia**, Rio de Janeiro, v. 60, p. 331-337, 2008.
- Dantas, M. E. Geomorfologia do estado do Rio de Janeiro. In: Silva, L. C.; Cunha, H. V. S. **Geologia do Estado do Rio de Janeiro: texto explicativo domapa geológico do Estado do Rio de Janeiro**. Brasília: CPRM, 2001.
- Escape, C. M.; Alemania, K. M.; Luzon, P. K.; Felix, R.; Salvosa, S.; Aquino, D.; Eco, R. N.; Lagmay, A. M. F. Comparison of various remote sensing classification methods for landslidedetection using ArcGIS. **Geophysical Research Abstracts**, v. 16: 15035, 2014.
- Francisco, C. N.; Almeida, C. M. Avaliação de desempenho de atributos estatísticos e texturais em uma classificação de cobertura da terra baseada em objeto. **Bol. Ciênc. Geod.**, v.18, n. 2, p. 302-326, 2012.
- Han, J.; Kamber, M. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers, 2006.
- Hölbling, D.; Friedl, B.; Eisank, C. An object-based approach for semi-automated landslide change detection and attribution of changes to landslide classes in northern Taiwan. **Earth Sci Inform**, 8:327–335, 2015.
- Klose, M.; Highland, L.; Damm, B., Terhorst, B. Estimation of direct landslide costs in industrialized countries: Challenges, concepts, and case study. In **Landslide Science for a Safer Geoenvironment**; Sassa, K., Canuti, P., Yin, Y., Eds.; Springer: Berlin, Germany, v. 2, p. 661–667, 2014.
- Körting, T. S. **GeoDMA: a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery**. 2012. 119 p. Tese (Doutorado em Sensoriamento Remoto) -Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012.
- Martha, T. R.; Norman K.; Cees J.; Van Westen, V. J.; Kumar, K. V. Object-Oriented Analysis of Multi-Temporal Panchromatic Images for Creation of Historical Landslide Inventories. **ISPRS Journal of Photogrammetry and Remote Sensing** 67 (2012): 105-19. Elsevier. Web. Sept. 2012.
- Netto, A. L. C.; Avelar, A. S.; D’Orsi, R. N. Vulnerabilidades dos sistemas naturais: monitoramento dos problemas de encosta na cidade do rio de janeiro frente às mudanças climáticas em curso e futuras. In: **Megacidades, vulnerabilidades e mudanças climáticas: Região metropolitana do Rio de Janeiro**, 2011. Disponível em: <http://www.poli.ufrj.br/noticias/arquivos/completo.pdf>. Acesso em: 29 jul. 2016.
- Quinlan, R. **C4.5: programs for machine learning**. San Francisco: Morgan Kaufmann. 1993.
- Robaina, L. E. S. Espaço Urbano: Relação com os acidentes e desastres naturais no Brasil. **Ciência e Natura**, v. 30, p. 107-126, 2008.
- Rodrigues, T. C. S. **Classificação da cobertura e do uso da terra com imagens WorldView-2 de setores norte da Ilha do Maranhão por meio do aplicativo InterIMAGE e de mineração de dados**. 2014. 87 p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.
- Selby, M.J. **Hillslope materials & processes**. New York: Oxford University Press, 1993. 451p
- Silva, M. P. S. et al. Mining patterns of change in remote sensing image databases. In: **IEEE Internacional Conference on data mining**, 15, 2005.
- UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC). Centro Universitário de Estudos e Pesquisas sobre Desastres. **Atlas Brasileiro de Desastres Naturais: 1991 a 2012 / Centro Universitário de Estudos e Pesquisas sobre Desastres**. 2. Ed. rev. Ampl. – Florianópolis: CEPED UFSC, 2013. 126 p.