BCG - Boletim de Ciências Geodésicas - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702015000400041

Artigo

MINERAÇÃO DE DADOS APLICADA À DISCRIMINAÇÃO DA COBERTURA DA TERRA EM IMAGEM LANDSAT 8 OLI

Data mining applied for land cover classification using Landsat 8

Guilherme Domingues dos Santos[1]

Cristiane Nunes Francisco[2]

Cláudia Maria de Almeida[3]

[1]Instituto Militar de Engenharia Pós-graduação em Engenharia Cartográfica.Praça General Tibúrcio, 80. Urca, Rio de Janeiro – RJ – 22290-27 guilhermedomsantos@gmail.com

[2]Universidade Federal Fluminense.Instituto de Geociências. Departamento de Análise Geoambiental. Campus da Praia Vermelha. Boa Viagem, Niterói - RJ − 24210-310.crisnf@vm.uff.br

[3]Instituto Nacional de Pesquisas Espaciais. Av. dos Astronautas, Caixa Postal 515 − 12245-970 São José dos Campos - SP, Brasil.almeida@dsr.inpe.br

Resumo:
O presente trabalho tem como objetivo investigar os descritores espectrais, extraídos das bandas do Landsat 8, e topográficos, provenientes de dados do TOPODATA, que auxiliam na discriminação das classes de cobertura da terra através de uma árvore de decisão gerada por mineração de dados. Foram extraídas medidas estatísticas de amostras referentes a 12 classes, coletadas no município do Rio de Janeiro, de um banco de dados composto por 18 planos de informação. Os resultados apontaram que entre os descritores estatísticos, prevaleceram a média e mediana. Como descritor espectral, merece destaque a banda 1 (ultra-azul), selecionada para discriminar, além das classes de água, as classes de vegetação e não-vegetação. O minerador utilizou o Índice de Vegetação por Razão Simples (RS) em todas as árvores a despeito do Índice de Vegetação por Diferença Normalizada (IVDN). A declividade também foi inserida nas três árvores para separar o afloramento rochoso da vegetação de baixo porte. Em relação aos níveis digitais, não foi utilizado nenhum descritor de radiância nas quatro árvores. Considerando o grande volume de dados produzidos e armazenados atualmente, pode-se afirmar que a mineração é um importante recurso para a extração de informações de volumosos bancos de dados em curto espaço de tempo.
Palavras-Chave: redes semânticas, classificação de imagens, árvores de decisão.

Abstract:
This paper is committed to investigate the spectral attributes extracted from Landsat 8 image bands and topographic attributes derived from TOPODATA, meant to discriminate land cover classes by means of decision trees, a technique in the scope of data mining. Statistical measures of samples corresponding to 12 land cover classes collected in Rio de Janeiro city were calculated from a database composed of 18 layers, from which four decision trees were generated. The results showed that the mean and median were the most relevant statistical attributes. As to spectral attributes, Band 1 is worth of mention, which has been selected to classify water classes, besides discriminating vegetation and non-vegetation classes. Regarding

vegetation indices, the data mining algorithm exclusively relied on the Simple Ratio Index in all trees to the detriment of the NDVI. Slope has been employed in three decision trees to separate rock outcrop from low-height vegetation. On the other hand, radiance has not been used in any of the four decision trees. Considering the ever-increasing volume of remotely sensed data currently available, it ought to be acknowledged that data mining represents a crucial solution to efficiently extract information from large databases in a short time.
Keywords: semantic networks, images classification, decision trees.

1. Introdução

Nas últimas décadas, com o avanço tecnológico na área de imageamento orbital, entraram em operação inúmeros satélites de sensoriamento remoto, aumentando a oferta de imagens da superfície terrestre e, consequentemente, a disponibilidade de dados com ampla gama de resoluções espaciais, espectrais e radiométricas. As imagens de sensoriamento remoto se constituem em uma fonte de informações sobre a cobertura da terra, que podem ser extraídas através de interpretação visual com o uso dos elementos contextuais de reconhecimento de padrões, por classificação automática pixel a pixel, ou por regiões. Esta última se subdivide em classificação por regiões convencional e baseada na análise de objeto (GEOBIA – *Geographic Object-based Image Analysis*), sendo que GEOBIA diferencia-se da classificação por regiões convencional por possuir um modelo de conhecimento atrelado ao processo de classificação (Hay; Castilla, 2008). No entanto, os métodos e técnicas de extração de informações precisam ser difundidos em maior escala, pois, apesar dos avanços nas resoluções das imagens de sensoriamento remoto, a grande maioria dos usuários ainda utiliza conceitos básicos e aplicativos de processamento de imagem desenvolvidos na década de 1970 (Blaschke; Strobl, 2001).

A interpretação visual de imagens orbitais de sensoriamento remoto utiliza os métodos de fotointerpretação, em especial, os elementos de reconhecimento para identificação do alvo, como forma, tonalidade, cor, localização, textura e padrão, em que o conhecimento e a experiência do analista a respeito da cena são fundamentais para resultados de boa qualidade, bem como as suas aptidões, como, por exemplo, o grau de percepção de detalhes (Jensen, 2009). Assim, o mapeamento pode ter diferentes resultados em função do perfil do intérprete. Além disso, o tamanho da área de estudo interfere diretamente no tempo de análise, uma vez que quanto maior a área, maior o tempo demandado para a execução da interpretação.

Para otimizar o tempo de trabalho e minimizar os erros provenientes da subjetividade da interpretação humana, as técnicas de classificação digital de imagens estão em constante aprimoramento, visando reduzir o esforço do analista e automatizar o processo de extração de informações. A classificação automática apresenta como resultado uma imagem digital formada por um conjunto de pixels ou objetos classificados que agregam os padrões homogêneos de alvos, dando origem a um mapa digital temático (Meneses; Almeida, 2012).

Os classificadores pixel a pixel baseiam-se na resposta espectral, ou seja, os elementos da imagem são analisados individualmente sem que o contexto espacial no qual estão inseridos seja considerado. No entanto, como os pixels vizinhos tendem a assumir características similares, maior tempo e esforço são necessários nos procedimentos de pós-classificação para atenuação de ruídos (Pinho, 2005).

Com a aplicação de GEOBIA, os objetos são criados antes da classificação propriamente dita através da segmentação da imagem, baseada no contexto onde os pixels estão inseridos, resultando no agrupamento de um conjunto de pixels com informações contextuais de cada

região, aumentando, assim, o número de elementos de reconhecimento que podem ser utilizados para discriminar uma dada classe de cobertura da terra. Esse tipo de classificação torna-se vantajoso, pois é possível alcançar resultados mais precisos em função da associação de cada região a um conjunto de atributos (descritores), e assim definir as características de cada objeto de acordo com suas propriedades contextuais, como forma, textura e topologia, indo além da resposta espectral (Francisco; Almeida, 2012a).

No entanto, a seleção dos descritores mais adequados para a discriminação das classes tem dificultado a execução da classificação por GEOBIA (Pinho, 2005), pois não há um padrão para discriminação de classes de cobertura da terra que determine os atributos que as melhor diferenciem, além do grande volume de dados espectrais que cresceu exponencialmente nas últimas décadas, aumentando consideravelmente a dificuldade na escolha dos atributos, bem como o tempo e esforços despendidos no trabalho. Para facilitar e aprimorar a seleção dos atributos mais adequados, atualmente estão sendo utilizados algoritmos de mineração de dados em estudos de classificação de imagens de sensoriamento remoto, que consistem em sistemas que realizam processos automáticos de extração de informações através do reconhecimento de padrões nos dados (Vieira, 2010).

A busca por padrões nos dados não é um recurso particularmente novo. Diversas áreas do conhecimento há algum tempo já trabalham com a ideia de que os padrões podem ser procurados automaticamente, identificados, validados e utilizados para a previsão. A nova questão está ligada ao aumento vertiginoso do volume de dados e das oportunidades para encontrar padrões (Witten et al., 2011). A velocidade do crescimento de uma enorme quantidade de dados coletados e armazenados em grandes e numerosos bancos de dados ultrapassou em muito a capacidade humana de compreensão dos mesmos, sem a existência de potentes ferramentas. Como resultado, os dados coletados em grandes bancos tornam-se "túmulos de dados", ou seja, arquivos-dados que raramente são visitados (Han; Kamber, 2006). Consequentemente, decisões importantes não são muitas vezes feitas com base em dados ricos em informações armazenadas, mas sim na intuição do tomador de decisão por não dispor de ferramentas para extrair o valioso conhecimento incorporado nas vastas quantidades de dados (Han; Kamber, 2006).

A mineração de dados é uma das ferramentas para análise de grandes volumes de dados, a qual integra uma área maior conhecida como Descoberta de Conhecimento a partir de Banco de Dados (DCBD) e consiste em um recurso automatizado para extração de informações a partir de dados previamente conhecidos (Witten et al., 2011). Entre as categorias de mineração de dados, encontra-se a árvore de decisão, aplicada a partir da formação de um banco de dados, que diferencia e reconhece padrões estatísticos a partir da aplicação de rotinas iterativas de testes para a divisão dos dados, tendo como resultado a geração de um modelo gráfico formado por nós (conceitos de classes, abstratas ou concretas) e arcos (conexões entre os nós), à semelhança de uma árvore. Assim, representada por um fluxograma, a árvore de decisão adota a estratégia *top-down* (de cima para baixo), na qual, a partir de um conjunto de amostras de dados, os atributos são discriminados em nós sucessivos até que se atinja um nível mais detalhado e, por fim, a classificação final do conjunto de dados (Vieira, 2010; Witten et al., 2011).

Como a mineração utiliza dados que podem ser expressos matematicamente, atributos estatísticos e contextuais dos objetos extraídos das imagens e de outras fontes, como os Modelos Digitais de Elevação (MDE), estes têm sido adotados para compor o banco de dados de entrada (Camargo, 2008). Os atributos estatísticos baseiam-se nas propriedades matemáticas das variações do número digital para descrever suas medidas, enquanto que os contextuais consideram a distribuição espacial dessas variações, como textura, forma e topologia. Assim, a mineração de dados, aplicada na classificação de imagens de sensoriamento remoto, tem sido utilizada para selecionar descritores que discriminam as classes de cobertura da terra, a partir de um banco de dados formado por atributos extraídos de um conjunto volumoso de bandas

multiespectrais e de imagens derivadas de processamento digital, além de outras fontes de dados (Leonardi, 2010).

Entre os sistemas orbitais em operação, o LANDSAT 8, lançado em 2013, merece destaque. Além de o Landsat ser o programa de satélites que obtém imagens por período mais longo e contínuo de tempo, desde 1972, quando foi lançado o Landsat 1, o sensor OLI, presente no satélite recentemente lançado, incorporou algumas inovações. A missão dos satélites Landsat representa uma das principais fontes de informações para a comunidade científica, por possuir uma série histórica de mais de 40 anos, abrigando um grande volume de dados relevantes da superfície da Terra disponibilizado gratuitamente no site do USGS (United States Geological Survey).

Das inovações incorporadas no LANDSAT 8/OLI, destaca-se a resolução radiométrica de 16 bits, que acarretou em ganho quali-quantitativo em função da maior precisão radiométrica registrada em comparação a imagens de outros sensores, que costumavam variar entre 8 e 11 bits. Em comparação aos sensores dos Landsat anteriores, o sensor OLI também apresenta uma redução nas larguras espectrais de todas as bandas, com destaque para a banda do infravermelho próximo (USGS, 2013). As bandas do visível sofreram as menores alterações, com amplitude entre 0 e 0,01 µm, enquanto as bandas do infravermelho a amplitude foi de 0,06 a 0,10 µm.

Incorporou, também, uma nova banda espectral denominada como ultra-azul (banda 1), situada na borda da faixa entre o ultravioleta e o azul, indicada para aplicação em estudos costeiros e sobre aerossóis. No entanto, as diferenças de resposta espectral entre os alvos "água", "solo" e "vegetação" tendem a ser realçadas nesse intervalo do espectro eletromagnético, diferenciando, além das classes de água entre si, as demais classes (solo, vegetação), e discriminando melhor as classes de vegetação mais densas, como a floresta, em que a resposta do solo é insignificante, daquelas menos densas, a exemplo de vegetação herbácea, em que a contribuição do solo é mais evidente. Uma indicação da diferenciação espectral entre esses alvos na banda ultra-azul é apresentada na Figura 1.

Nesse contexto, o presente trabalho tem como objetivo investigar os descritores espectrais extraídos das bandas do Landsat 8, e topográficos, provenientes de dados do TOPODATA, que auxiliam na discriminação das classes de cobertura da terra, representadas em árvores de decisão geradas por meio de um sistema de mineração de dados.

As amostras foram coletadas no município do Rio de Janeiro, com extensão de 1,2 mil km², por apresentar uma diversidade de classes de cobertura da terra. A área de estudo é caracterizada por apresentar relevo plano entremeado por maciços cristalinos e morros. A classe de cobertura da terra predominante é a urbana, localizada principalmente nas áreas planas; no entanto, nas encostas e morros da cidade, é também comum a presença do uso urbano, mas com grande variedade na densidade de ocupação, sendo que as de muito baixa densidade correspondem às áreas recentes de expansão urbana. No interior dos maciços, predominam as florestas em diversos estágios de sucessão, contudo, também é comum a presença de afloramentos rochosos e áreas cobertas por campo antrópico. Por estarem localizados na área litorânea, embora distribuídos em pequenas manchas, também estão presentes a vegetação de restinga e os manguezais, além de praias, dunas, costões rochosos, baías e lagoas costeiras.

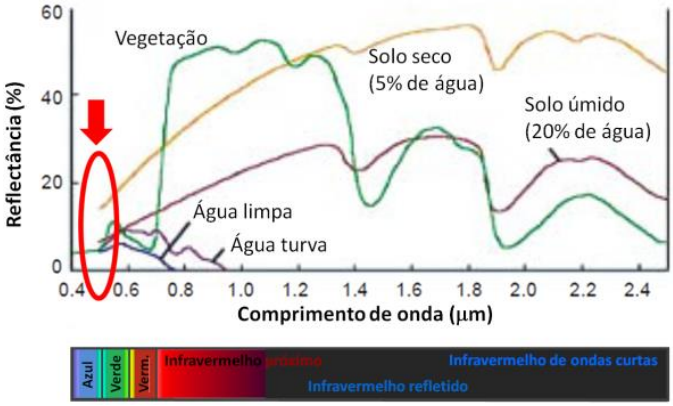


Figura 1: Comportamento espectral de alvos em relação ao espectro eletromagnético nas faixas do visível, infravermelho próximo e de ondas curtas. Fonte: Jiang (2013).

2. Material e métodos

O banco de dados para mineração foi composto por sete bandas do sistema Landsat 8 OLI, obtidas no site do USGS, referentes a um recorte de cena gerada em 14 de maio de 2013: a banda 1 correspondendo ao ultra-azul; as bandas 2, 3 e 4, correspondendo, respectivamente às regiões do azul, verde e vermelho; a banda 5 correspondendo ao infravermelho próximo e, por fim, as bandas 6 e 7 correspondendo à faixa espectral do infravermelho de ondas curtas.

Os números digitais foram convertidos para radiância visando realizar a caracterização espectral das classes, bem como executar operações aritméticas entre as bandas espectrais com a finalidade de gerar os índices de vegetação. Como os números digitais possuem um intervalo dado pela resolução radiométrica cujos valores extremos são associados ao maior e ao menor valores de radiância registrados pelo sensor em cada faixa espectral, recomenda-se que sejam convertidos para radiância nas operações em que há o processamento conjunto de bandas. A não-conversão para valores físicos acarreta erros nos índices, pois os números digitais não correspondem à mesma escala radiométrica em todas as bandas (Ponzoni; Shimabukuro, 2009).

A conversão para a radiância das bandas do LANDSAT 8 foi feita com base na equação de calibração (Equação 1) obtida em Chander et al. (2009), tendo sido aplicados os valores de radiância mínima e máxima disponíveis no arquivo de metadados que acompanha a cena (Tabela 1).

$$L_0 = L_{min} + \left(\frac{\left(L_{max}(\lambda) - L_{min}(\lambda)\right)}{2^x} * ND(\lambda) \right) \quad (1)$$

Em que:
L_o = radiância aparente;
$L_{min}(\lambda)$ = radiância mínima;
$L_{max}(\lambda)$ = radiância máxima;
ND = número digital; e
X = número de bits da imagem.

Tabela 1: Radiância das bandas do Landsat 8 obtidas em 14/05/2013.

Bandas	Radiância Máxima	Radiância Mínima
1	764,55450	-63,13716
2	779,64441	-64,38329
3	713,89124	-58,95337
4	604,62219	-49,92990
5	366,89148	-30,29802
6	92,43834	-7,63359
7	30,07002	-2,48319

A seguir, foram calculados os índices de vegetação IVDN (Índice de Vegetação por Diferença Normalizada) e RS (Razão Simples). De acordo com Jensen (2009), os dois índices apresentam desempenhos diferentes de acordo com o tipo de vegetação. Enquanto o IVDN tem melhor desempenho nas classes de vegetação de menor biomassa, como pastos, áreas áridas e semi-áridas, saturando nas classes de maior biomassa, como florestas e áreas de vegetação densa, o desempenho do RS é o contrário, ou seja, apresenta maior intervalo com vegetação de maior biomassa, saturando na vegetação de menor biomassa.

Também foram incorporados ao banco de dados o MDE e a declividade oriundos do TOPODATA, que consiste em um banco de dados de variáveis geomorfométricas para todo o território nacional, com resolução espacial de 30 m, derivadas do processamento dos modelos da Missão de Topografia por Radar do Ônibus Espacial (*Shuttle Radar Topography Mission* - SRTM) (Valeriano, 2005).

O banco de dados final foi constituído por 18 planos de informação: sete bandas com número digital, sete bandas com valores de radiância, duas imagens com índices de vegetação (RS e IVDN), um MDE e uma grade de declividade. Optou-se por manter as bandas em números digitais e em radiância, objetivando compará-los na discriminação das classes de cobertura da terra.

Posteriormente, por interpretação visual, foram definidas 12 classes de cobertura da terra presentes na área de estudo: água limpa, água turva, afloramento rochoso, apicum, área urbana, brejo, floresta, herbácea, lagoa, mangue, solo arenoso, solo argilo-arenoso (Quadro 1). Para cada classe, foram coletadas no mínimo 30 amostras com área de aproximadamente 90 mil m², para que cada uma contivesse cerca de 100 pixels (Figura 2).

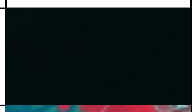
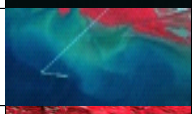
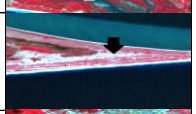
A partir da sobreposição entre 373 amostras e 18 planos de informação, foram extraídas medidas estatísticas dos números digitais de cada amostra - média, mediana e desvio-padrão, a partir do programa ArcGIS Desktop V. 10, e gerado um banco de dados contendo as amostras (373 registros) e seus respectivos atributos (54 campos).

Os atributos estatísticos média e mediana representam medidas de tendência central dos números digitais das amostras, enquanto o desvio-padrão é uma medida de dispersão dos números digitais e, como representa a variabilidade de um conjunto de dados, pode ser utilizado como uma medida de textura na classificação de imagens, requerendo menor processamento do que as matrizes de co-ocorrência (Baraldi; Panniggiani, 1995). Enquanto as matrizes baseiam-se na distribuição espacial dos números digitais para o cálculo de textura, o desvio-padrão, considerado como uma medida estatística de textura de primeira ordem, é calculado com base no histograma dos números digitais de cada pixel (Maji; Pal, 2008). A relação entre textura e desvio-padrão é corroborada no trabalho de Francisco & Almeida (2012b) cuja árvore de decisão, gerada apenas por atributos estatísticos, apresentou o desvio-padrão para discriminar a área urbana de todas as outras classes investigadas já no primeiro ramo da árvore.

A seguir, com o auxílio de um editor de texto, o banco de dados foi convertido para o formato ARFF (*Attribute-Relation File Format*) e importado para o programa WEKA (*Waikato Environment for Knowledge Analysis*), sistema de licença livre, emitida pela GNU *General Public License*, e que utiliza um conjunto de algoritmos de aprendizado de máquina para executar a mineração dados.

As regras de classificação por mineração de dados utilizadas neste trabalho foram estabelecidas pelo algoritmo C4.5, criado por Quinlan (1993) e implementado como classificador tree.J48 no programa Weka 3.6.4. Esse algoritmo constrói árvores de decisão a partir de amostras de treinamento, que, conforme anteriormente exposto, consistem em um fluxograma no qual o nó interno representa um teste com um atributo (Figura 3), correspondendo a uma classe abstrata; o ramo corresponde ao resultado do teste, e a classe esperada ou concreta é exibida pelo nó externo ou nó-folha. As classes são separadas de acordo com o atributo que o algoritmo julga ser o mais adequado para discriminá-las, sendo irrelevantes os descritores que não são utilizados pela árvore (Francisco; Almeida, 2012b).

Quadro 1: Classes de cobertura da terra.

Classes	Definição	Chave de interpretação	Landsat 8 R(5) G(4) B(3)
Água limpa	Corpo d'água não confinado sem presença de sedimentos em suspensão.	• Cor – azul/preto • Textura – lisa •Tom – escuro	
Água turva	Corpo d'água não confinado com sedimentos em suspensão.	• Cor – azul/ciano/marrom • Textura – lisa a intermediária •Tom – claro	
Afloramento rochoso	Rocha exposta.	• Cor – marrom • Textura – Intermediária •Tom – intermediário	
Apicum	Área arenosa de teor salino alto e levemente alagada.	• Cor – ciano • Textura – lisa •Tom – claro	
Área urbana	Área com presença predominante de construções.	• Cor – ciano • Textura – rugosa •Tom – claro	
Brejo	Área alagadiça com presença de vegetação de baixo porte.	• Cor – marrom • Textura – lisa a intermediária •Tom – intermediário	
Floresta	Vegetação densa de alto porte.	• Cor – vermelho • Textura – rugosa •Tom – claro	
Herbácea	Predominantemente constituído por vegetação de baixo porte.	• Cor – vermelho • Textura – lisa a intermediária •Tom – intermediário	
Lagoa	Corpo d'água confinado.	• Cor – azul • Textura – lisa •Tom – escuro	
Mangue	Ambiente de transição entre terra e mar, alagadiço e/ou lamoso, com vegetação de porte médio com raízes aéreas	• Cor – vermelho/marrom • Textura – intermediária •Tom – intermediário	
Solo arenoso	Solo exposto predominantemente arenoso.	• Cor – branco • Textura – intermediária •Tom – claro	
Solo argilo-arenoso	Solo exposto argilo-arenoso.	• Cor – ciano • Textura – lisa a intermediária •Tom – claro (tonalidade mais clara que do arenoso)	

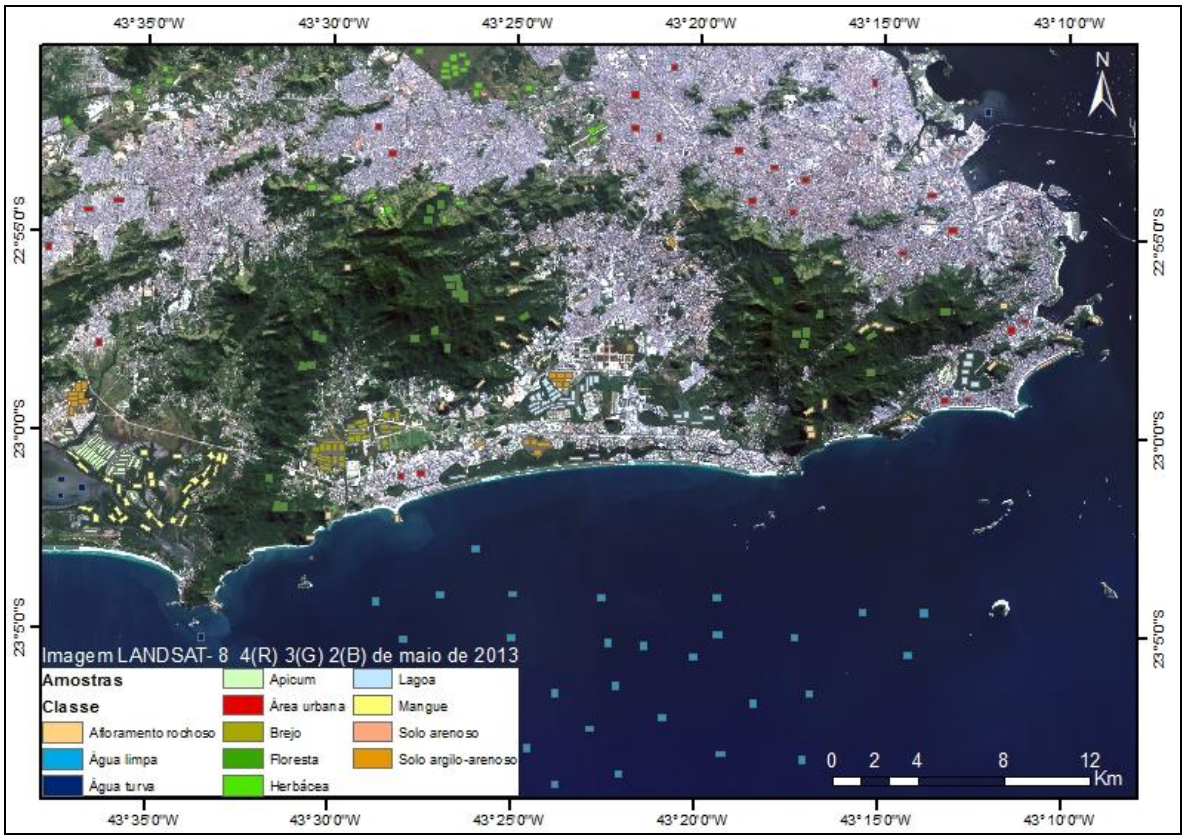


Figura 2: Amostras de treinamento coletadas na cidade do Rio de Janeiro**.**

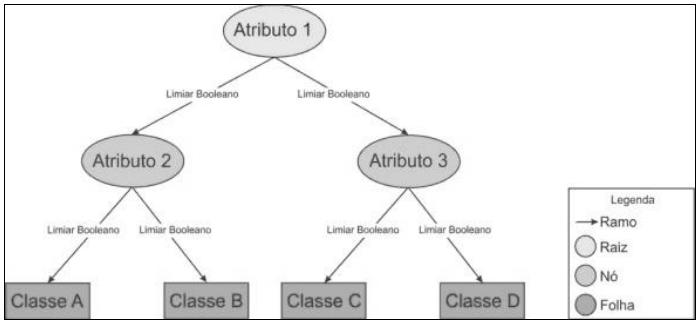


Figura 3: Exemplo de árvore de decisão.

3. Resultados

Foram construídas quatro árvores de decisão, buscando aquela que fosse menos complexa e pouco ramificada, para que pudesse ser reproduzida em outros estudos. A primeira (Figura 4) foi gerada sem limitação do número de amostras por descritor, o que facilitou a criação de nós para discriminação de poucas amostras e, assim, vários descritores foram selecionados, dispostos em diferentes posições na árvore, para discriminar um pequeno número de amostras por descritor. Com isso, o resultado foi uma árvore muito ramificada e complexa cujos índices de acurácia, obtidos nos testes executados pelo minerador com as amostras de treinamento, resultaram em índice Kappa de 0,91 e de exatidão global de 92%, enquanto a exatidão do produtor ou do consumidor ficaram abaixo de 90% apenas para as classes de solo, afloramento rochoso e vegetação herbácea (Tabela 2).

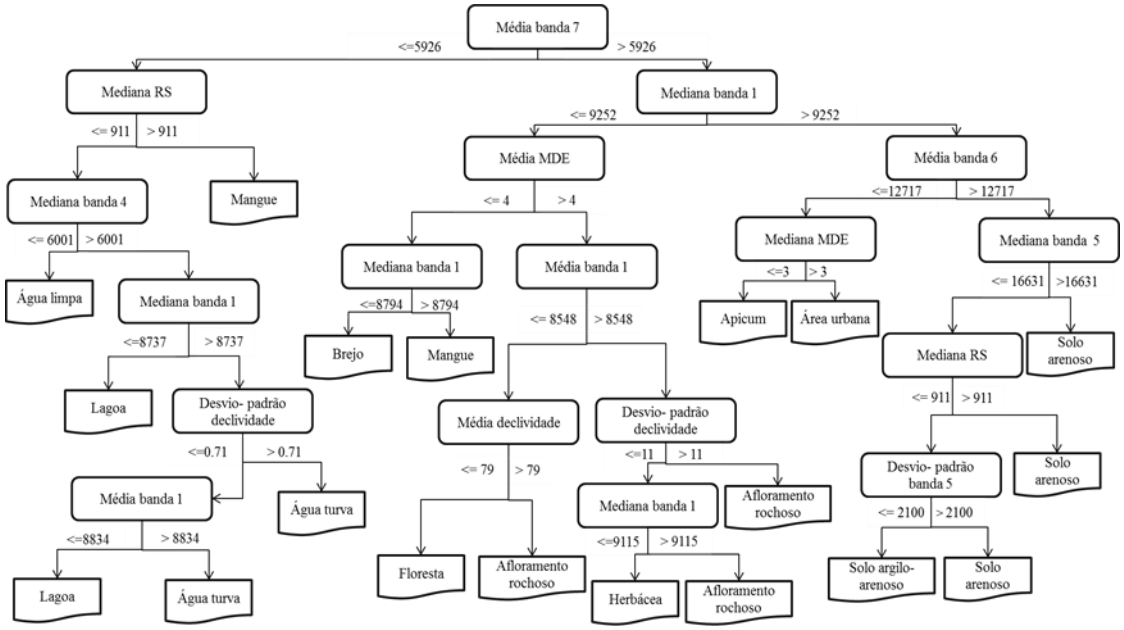


Figura 4: Árvore de decisão gerada sem limite de amostras por descritor.

Tabela 2: Matriz de confusão da primeira árvore de decisão.

	Classes	Amostras de referência												
		Floresta	Mangue	Solo aren.	Afl. rochoso	Apicum	Brejo	Solo arg-aren.	Herbácea	Lagoa	Água limpa	Água turva	Área urbana	Total classifica-das
Classificadas	Floresta	29			3									32
	Mangue		29				1							30
	Solo arenoso			28	1			4						33
	Afloramento rochoso				24				2					26
	Apicum					30								30
	Brejo		1				29	1						31
	Solo argilo-arenoso			2				23	2				1	28
	Herbácea	1	1				1		26					29
	Lagoa									31		1		32
	Água limpa				1						32			33
	Água turva									3	1	31		35
	Área urbana				1			2					31	34
Total de amostras		30	31	30	30	30	31	30	30	34	33	32	32	373
Exatidão do produtor		0,97	0,94	0,93	0,80	1,00	0,94	0,77	0,87	0,91	0,97	0,97	0,97	0,92
Exatidão do consumidor		0,91	0,97	0,85	0,92	1,00	0,94	0,82	0,90	0,97	0,97	0,97	0,91	

A segunda árvore (Figura 5) foi gerada com a limitação de 15 amostras por cada descritor selecionado, resultando em uma árvore com menos nós e, portanto, menos ramificada e complexa. Assim como na primeira, os índices de acurácia, calculados pelo minerador com as amostras de treinamento, apresentaram-se elevados, com índice Kappa correspondendo a 0,90 e de exatidão global a 91%, enquanto a exatidão do produtor e do consumidor ficaram abaixo de 90% apenas para as classes de solo, afloramento rochoso e vegetação herbácea (Tabela 3). No entanto, o descritor MDE foi selecionado para discriminar classes que podem ocorrer em mesmas altitudes, como as classes área urbana e apicum. Em outro ramo, a classe brejo foi discriminada das classes afloramento rochoso, herbácea e floresta também pelo descritor MDE

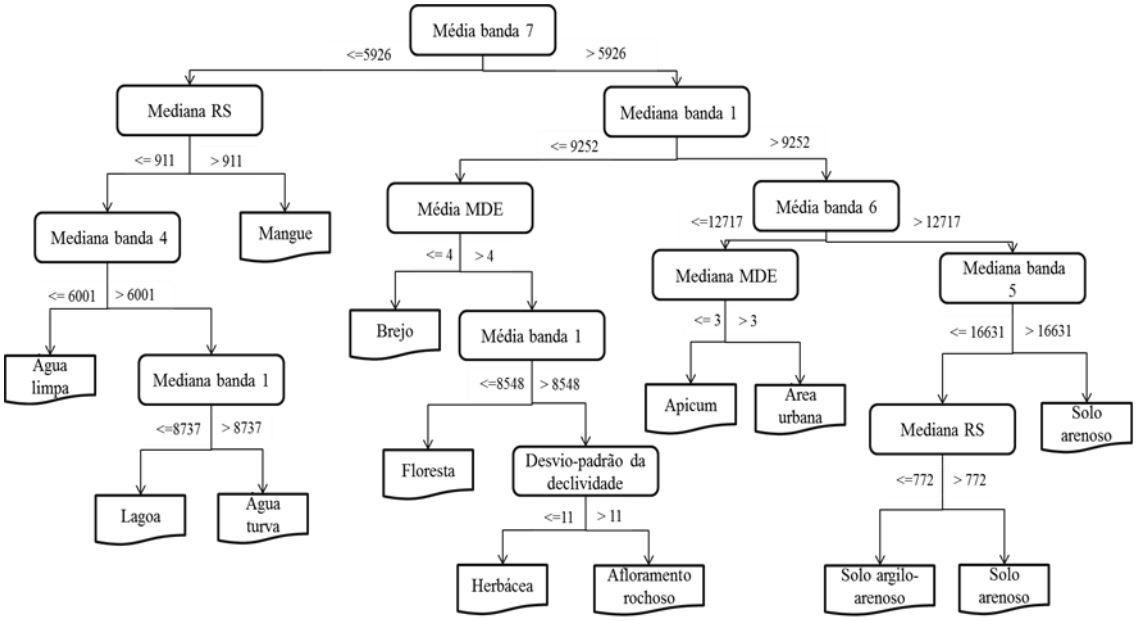


Figura 5: Árvore de decisão gerada com no mínimo 15 amostras por descritor.

Tabela 3: Matriz de confusão da segunda árvore de decisão.

	Classes	Amostras de referência												
		Floresta	Mangue	Solo aren.	Afl. rochoso	Apicum	Brejo	Solo arg-aren.	Herbácea	Lagoa	Água limpa	Água turva	Área urbana	Total classifica-das
Classificadas	Floresta	29			3									32
	Mangue		29											29
	Solo arenoso			23				5						28
	Afloramento rochoso				24				3					27
	Apicum					30								30
	Brejo		1				30	1	1					33
	Solo argilo-arenoso			7	1			22					1	31
	Herbácea	1	1				1		26					29
	Lagoa									31				31
	Água limpa				1						32			33
	Água turva									3	1	32		36
	Área urbana				1			2					31	34
Total de amostras		30	31	30	30	30	31	30	30	34	33	32	32	373
Exatidão do produtor		0,97	0,94	0,77	0,80	1,00	0,97	0,73	0,87	0,91	0,97	1,00	0,97	0,91
Exatidão do consumidor		0,91	1,00	0,82	0,89	1,00	0,91	0,71	0,90	1,00	0,97	1,00	0,91	

A terceira árvore (Figura 6) foi gerada com limitação de 30 amostras por cada descritor selecionado e sem o descritor MDE, a título de investigação. Foi verificada elevada mistura entre as classes na matriz de confusão, elaborada pelo minerador com base nas amostras de treinamento, o que justifica índice Kappa de apenas 0,60 e de exatidão global de 64%, enquanto predominaram índices de exatidão do produtor e do consumidor abaixo de 80%, com destaque para as classes herbácea e solo argilo-arenoso que não foram classificadas e, assim, não estão representadas na árvore de decisão (Tabela 4).

Uma quarta árvore (Figura 7) foi gerada com limitação de 15 amostras por cada descritor selecionado e com a exclusão do descritor MDE, resultando em uma árvore pouco ramificada, com índice Kappa de 0,89 e de exatidão global de 90%, enquanto a exatidão do produtor e consumidor ficaram abaixo de 80% apenas nas classes de solo exposto (Tabela 5).

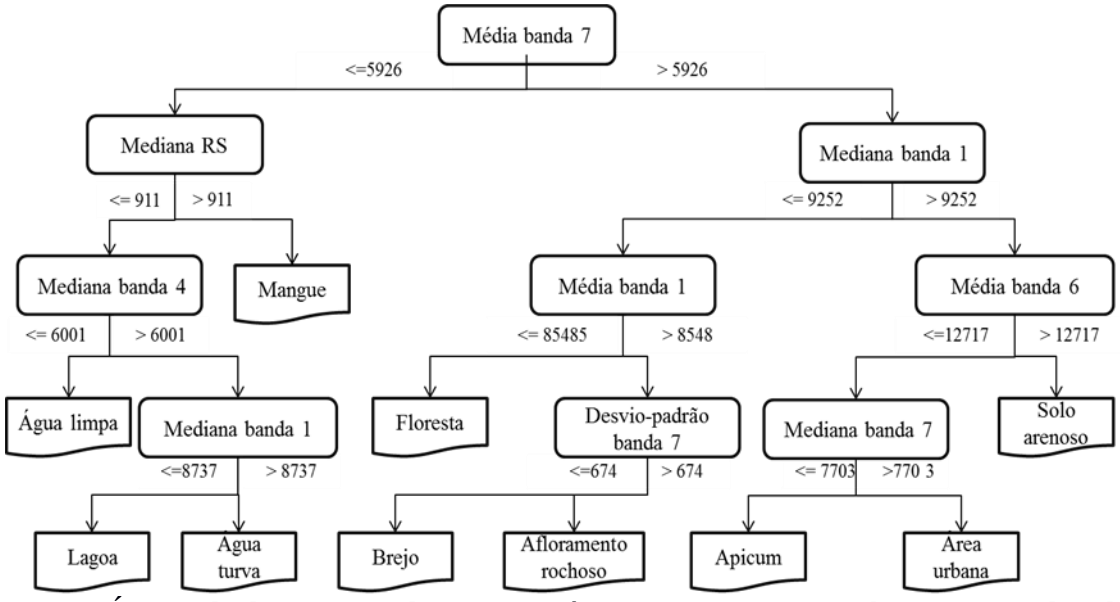


Figura 6: Árvore de decisão gerada com no mínimo 30 amostras por descritor e excluindo o MDE.

Tabela 4: Matriz de confusão da terceira árvore de decisão.

	Classes	Amostras de referência												
		Floresta	Mangue	Solo aren.	Afl. rochoso	Apicum	Brejo	Solo arg-aren.	Herbácea	Lagoa	Água limpa	Água turva	Área urbana	Total classifica-das
Classificadas	Floresta	23			3									26
	Mangue	6	23		1							1		31
	Solo arenoso			30	1			27					1	59
	Afloramento rochoso				13				5					18
	Apicum													0
	Brejo	1	1		1		31		11					45
	Solo argilo-arenoso													0
	Herbácea				9			1	14					24
	Lagoa		7							32		22		61
	Água limpa				1						32			33
	Água turva									2	1	9		12
	Área urbana				1	30		2					31	64
Total de amostras		30	31	30	30	30	31	30	30	34	33	32	32	373
Exatidão do produtor		0,77	0,74	1,00	0,43	0,00	1,00	0,00	0,47	0,94	0,97	0,28	0,97	0,64
Exatidão do consumidor		0,88	0,74	0,51	0,72	-	0,69	-	0,58	0,52	0,97	0,28	0,48	

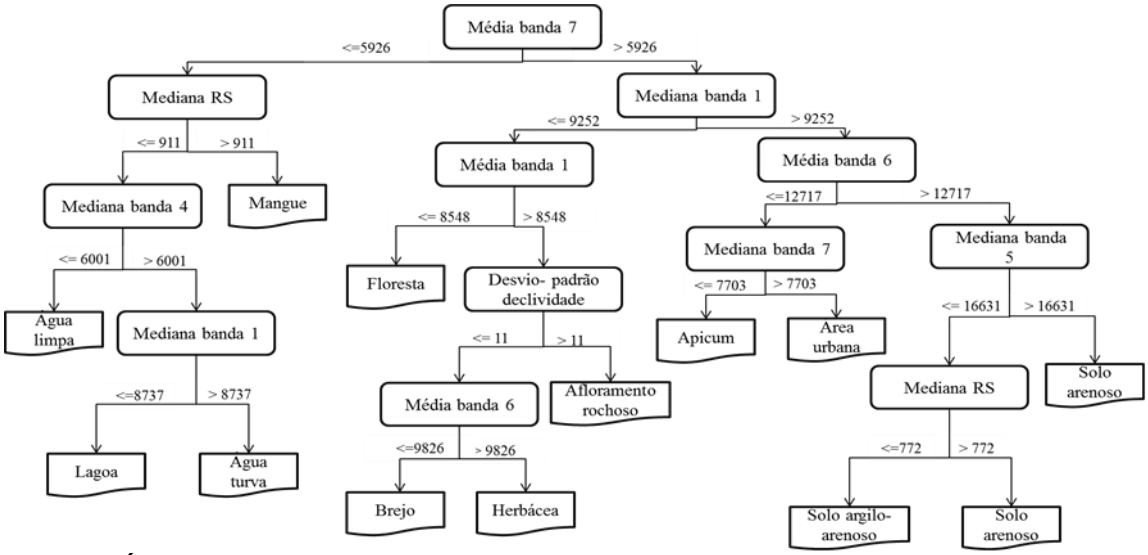


Figura 7: Árvore de decisão gerada com no mínimo 15 amostras por descritor e excluindo o MDE.

Tabela 5: Matriz de confusão da quarta árvore de decisão.

	Classes	Amostras de referência												
		Floresta	Mangue	Solo aren.	Afl. rochoso	Apicum	Brejo	Solo arg-aren.	Herbácea	Lagoa	Água limpa	Água turva	Área urbana	Total classifica-das
Classificadas	Floresta	29			3									32
	Mangue		29				1							30
	Solo arenoso			28	1			4						33
	Afloramento rochoso				24				2					26
	Apicum					30								30
	Brejo		1				29	1						31
	Solo argilo-arenoso			2				23	2				1	28
	Herbácea	1	1				1		26					29
	Lagoa									31		1		32
	Água limpa				1						32			33
	Água turva									3	1	31		35
	Área urbana				1			2					31	34
Total de amostras		30	31	30	30	30	31	30	30	34	33	32	32	373
Exatidão do produtor		0,97	0,94	0,93	0,80	1,00	0,94	0,77	0,87	0,91	0,97	0,97	0,97	0,92
Exatidão do consumidor		0,91	0,97	0,85	0,92	1,00	0,94	0,82	0,90	0,97	0,97	0,97	0,91	

Para formação dessa árvore, formada por 12 nós, estavam disponíveis 54 descritores, mas somente nove foram utilizados pelo minerador, referentes a médias ou medianas dos objetos em número digital nas bandas 1, 4, 5, 6 e 7, mediana do índice de vegetação RS e o desvio-padrão da declividade. Na matriz de confusão desta árvore, não foram identificados erros relevantes que interferissem na discriminação das classes, com exceção das classes de solo exposto, as quais apresentaram confusão entre si.

Inicialmente, a árvore de decisão discrimina, a partir da média da banda do infravermelho de ondas curtas (banda 7), em número digital, as classes que contêm água das que não contêm água, dividindo-as em dois grandes ramos. No ramo da esquerda, no segundo nível da árvore, foi utilizada a mediana da RS para separar mangue das classes puras de água. Em seguida, a mediana da banda 4, vermelho, em número digital, foi usada para separar água limpa, e a mediana da banda 1, ultra-azul, em número digital, para discriminar água turva e lagoa. No ramo da direita, no segundo nível da árvore, foi usada a mediana da banda 1, em número digital, para separar vegetação de não-vegetação, criando dois novos ramos. No ramo da esquerda, no terceiro nível, a média da banda 1, em número digital, foi utilizada para separar floresta, e, em seguida, foi usado o desvio-padrão da declividade para isolar afloramento rochoso da vegetação de baixo porte, que, por sua vez, foi diferenciada entre herbácea e brejo pela média da banda 6, infravermelho de ondas curtas, em número digital. No ramo da direita, no terceiro nível, a média da banda 6, em número digital, separou as classes de solo da área urbana e do apicum, sendo que os dois últimos foram separados pela mediana da banda 7, em número digital. As classes de solo exposto, solo arenoso e solo argilo-arenoso, foram discriminadas pela mediana da banda 5, infravermelho próximo, em número digital, e, em seguida, pela mediana da RS.

4. Conclusões

O banco de dados deste trabalho, formado a partir do recorte de cena Landsat 8, contendo as bandas originais e demais imagens provenientes de processamentos digitais deste além dos dados do TOPODATA, permitiu a geração de quatro árvores de decisão aqui apresentadas, que apresentaram Kappa entre 0,60 e 0,91. Através dos testes realizados, investigou-se o comportamento das árvores ao se retirar o MDE da estatística das amostras, uma vez que ele foi selecionado para discriminar classes situadas em altitudes próximas, como apicum de área urbana, e afloramento rochoso de vegetação herbácea e floresta. Também foi constatada a

importância da redução do número de amostras por descritor, o que possibilita criar árvores menos complexas e ramificadas. No entanto, se o número de amostras for alto, algumas classes podem não ser inseridas na árvore.

Em relação aos descritores estatísticos, prevaleceram a média e mediana. O desvio-padrão deve mais bem investigado, possivelmente em experimentos com imagens de diferentes resoluções espaciais, bem como com o uso de amostras que apresentem área de maior extensão. Neste trabalho as amostras apresentaram cerca de 100 pixels, de modo que possuem número de pixels suficiente para melhor representar a variabilidade da resposta espectral dos alvos.

Como descritor espectral, vale enfatizar a eficácia da banda 1, nova faixa espectral do Landsat 8, para discriminar diferentes tipos de cobertura da terra, não ficando restrita somente à identificação de classes de água, conforme indicado na própria documentação do Landsat 8 disponível no site do USGS. Essa banda foi selecionada para discriminar, além das classes de água, as classes de vegetação de não-vegetação, e diferenciar classes de vegetação. Resultados semelhantes foram alcançados por Francisco & Almeida (2012b), em trabalho sobre cobertura da terra com imagens ALOS/AVNIR utilizando mineração de dados e GEOBIA, no qual o minerador selecionou a banda 1, correspondente à faixa espectral entre 0,42 a 0,50 µm, também para discriminar floresta da vegetação herbácea. Consideram-se os resultados como uma importante contribuição para os estudos de classificação de imagens, pois as bandas do azul são tradicionalmente consideradas eficientes na discriminação de alvos aquáticos e, algumas vezes desprezadas, pela interferência atmosférica que sofrem.

Também merecem destaque as bandas do infravermelho, em especial, a banda 7 utilizada para separar o ramo contendo as classes de vegetação do ramo com as classes de não-vegetação em todas as árvores geradas.

Em relação ao índice de vegetação, o minerador utilizou o RS em todas as árvores, enquanto o IVDN não inserido em único ramo, o que pode ser justificado pelo fato de o RS fornecer resultados mais abrangentes para florestas e de variação reduzida para regiões de menor biomassa (Jensen, 2009).

Em relação aos níveis digitais, o minerador não selecionou nenhum descritor de radiância nas quatro árvores geradas, o que indica que essa grandeza física, apesar de ser fundamental para a geração coerente de índices de vegetação, não apresenta relevância para a discriminação das classes quando comparada aos números digitais.

Por fim, dever ser enfatizada a inserção da declividade em três árvores para separar o afloramento rochoso da vegetação de baixo porte, conforme ocorrido no trabalho de Francisco e Almeida (2012b), enfatizando assim a eficácia do descritor "declividade" para esse tipo de discriminação.

Considerando o grande volume de dados produzidos e armazenados atualmente, pode-se afirmar que a mineração é um importante recurso para a extração de informações de volumosos bancos de dados em curto espaço de tempo. Deve ser ressaltada também que a classificação de imagem, elaborada a partir de uma árvore de decisão produzida por meio de mineração de dados, pode apresentar-se vantajosa, pois permite que a seleção dos descritores seja feita de forma automática, não sendo influenciada pela subjetividade do intérprete. Além disso, a rede semântica obtida pode ser replicada em outras pesquisas com características similares de distribuição espacial e cobertura da terra (Camargo et al., 2009).

REFERÊNCIAS BIBLIOGRÁFICAS

Baraldi, A., and Parmiggiani, F. .1995. An Investigation of the Textural Characteristics Associated with Gray Level Co-occurrence Matrix Statistical Parameters. *IEEE Transactions on Geoscience and Remote Sensing* 33, no. 2: 293-304.

Blaschke, T., and Strobl, J. . 2001. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. *Remote Sensing Classification for Social Science*: 12- 17.

Camargo, F. F. 2008. *Análise Orientada a Objeto aplicada ao mapeamento de unidades morfológicas a partir de dados ASTER/TERRA* – SP, Masters diss., Instituto Nacional de Pesquisas Espaciais.

Camargo, F. F., Florenzano, T. G., Almeida, C. M., and Oliveira, C. O. . 2009. Geomorphological Mapping Using Object-Based Analysis and ASTER DEM in the Paraíba do Sul Valley, Brazil. *International Journal of Remote Sensing* 30:6613-6620.

Chander, G., Markham, B. L., and Helder, D. L. . 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sensing of Environment*, 113: 893–903.

Francisco, C. N., and Almeida, C. M. . 2012. Interpretação de imagens orbitais por meio de sistema especialista para o mapeamento de cobertura da terra em região montonhosa. *Revista Sociedade & Natureza* 24, no. 2: 283-302.

Francisco, C. N., and Almeida, C. M. . 2012. Avaliação de desempenho de atributos estatísticos e texturais em uma classificação de cobertura da terra baseada em objeto. *Boletim de Ciências Geodésicas* 18, no. 2: 302- 326.

Han, J., and Kamber, M. . 2006. Data mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

Hay, G. J., and Castilla, G. . 2008. Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline? In: Blaschke, Thomas, Stefan Lang, and Geoffrey J. Hay (Eds.) *Object-based image analysis spatial concepts for knowledge-driven remote sensing applications*. Springer: 75-89.

Jensen, J. R. . 2009. *Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres*. Ed. José Carlos Neves Ephiphanio. São José do Campos: Parêntese.

Jiang, C. Y. H. . 2013. Digital Elevation Model and Satellite Imagery Based Bushfire Simulation. *American Journalof Geographic Information System* 2, no. 4: 108-120.

Leonardi, F. . 2010. *Abordagens cognitivas e mineração de dados aplicadas a dados ópticos orbitais e de laser para a classificação de cobertura do solo urbano – SP*, Mestrado diss., Instituto Nacional de Pesquisas Espaciais.

Maji P., and Pal, S. K. . 2008. Maximum class separability for rough-fuzzy C-means based brain MR image segmentation. In: Peters, J. F.; Skowron, A. (Eds.) *Transactions on rough sets* IX. Springer: 114-134.

Meneses, P. R., and Almeida, T. . 2012. *Introdução ao Processamento de Imagens de Sensoriamento Remoto*. Brasília: Universidade de Brasília.

Pinho, C. M. D. . 2005. *Análise orientada a objetos de imagens de satélite de alta resolução espacial aplicada à classificação de cobertura do solo no espaço intraurbano: o caso de São José dos Campos – SP*, Masters diss., Instituto Nacional de Pesquisas Espaciais.

Ponzoni, F. J., and Shimabukuro, Y. E. . 2009. *Sensoriamento remoto no estudo da vegetação*. São José dos Campos: Parêntese.

Quinlan, R. . 1993. *C4.5: programs for machine learning.* San Francisco: Morgan Kaufmann.

Valeriano, M. M. . 2005. Modelo digital de variáveis morfométricas com dados SRTM para o território nacional: o projeto TOPODATA. Paper presented at the XII Simpósio Brasileiro de Sensoriamento Remoto, April 16-21, in Goiânia, Brasil.

Vieira, M. A. . 2010. Análise de imagem orientada a objeto e mineração de dados aplicadas ao mapeamento da cultura da cana-de-açúcar, Masters diss. Instituto Nacional de Pesquisas Espaciais.

Witten, I. H., Frank, E., and Hall, M. A. . 2011. *Data mining: practical machine learning tools and techniques*. Burlington: Morgan Kaufmann Publishers.

USGS. United States Geological Survey. Landsat Missions. http://landsat.usgs.gov/. (accessed 15 julho 2013).

Recebido em Novembro de 2014.

Aceito em Setembro de 2015.