

Mineração de dados para análise da cobertura florestal

Nathália Suemi Saito¹,
Fernanda Viana Paiva Arguello¹,
Maurício Alves Moreira¹

¹ Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{nssaito; fernanda.arguello; mauricio}@dsr.inpe.br

Abstract. The landscape ecology metrics associated with spectral and space data mining can be used to increase the potential of the analysis and applications of remote sensing data becoming an important tool for decision making. The present study sought to use data mining techniques and metrics of landscape ecology to classify and quantify the different types of vegetation present in São Luis do Paraitinga city, Sao Paulo, Brazil. The images went through object-oriented analysis through the plugin GeoDMA to obtain spectral and spatial data. This information was used to classify classes by decision trees. Eucalyptus and Forest fragment areas represented 8.6% and 36,1% of the total area, respectively. The decision tree generated by the classification algorithm was used to obtain the map of forest cover. The classification by decision tree showed kappa of 0.80, indicating little confusion. The results indicate the importance of the forest sector and contribute in studies to contain the impacts and problems caused by the expansion of eucalyptus plantations in the municipality. The generation of classifications by the method of data mining metrics associated with landscape ecology proved to be an affordable and reliable tool to extract the spatial and spectral data with remote sensing techniques. The method was efficient for the separation of forest classes and the metrics used were important to better understand the objects inserted in the landscape and the pressures they suffer.

Palavras-chave: data mining, forest cover, interpretação de imagens, GeoDMA, eucalipto

1.Introdução

A vegetação representa um importante indicador das condições ambientais e contribui para manter a biodiversidade de uma região por proteger o solo, reduzir o transporte de sedimentos e o assoreamento dos cursos de água, além de servir de habitat para animais silvestres. Entretanto, nas últimas décadas o processo de transformação da paisagem natural em uma paisagem antropizada intensificou a pressão sobre as áreas naturais, o que resulta em ambientes pouco diversificados, com fragmentos florestais isolados e de dimensões reduzidas (CALEGARI et al., 2010).

No Brasil, o bioma da Mata Atlântica, com apenas 11,8% de sua cobertura original, possui uma longa história no uso da terra, importante na compreensão dos processos resultantes da transformação dessas áreas (RIBEIRO et al., 2009; FREITAS et al., 2010) que vão desde eventos como enchentes, inundações até deslizamentos de terra.

Diante da problemática causada pelas alterações da cobertura vegetal, torna-se importante conhecer esta cobertura, bem como a estrutura da paisagem para auxiliar no monitoramento e planejamento do ambiente (SAITO et al., 2010). Para isso, utilizam-se dados de sensoriamento remoto, como por exemplo, as imagens orbitais, para detectar, mapear e quantificar os alvos da superfície terrestre.

Entretanto, segundo Moreira (2011), deve-se tomar certos cuidados ao serem utilizadas imagens de média a baixa resolução espacial para interpretar alvos na superfície terrestre. Pois, nestas imagens, na maioria das vezes, não é possível individualizar um único alvo, sendo a interpretação feita de forma dedutiva, em que o alvo é identificado através da associação das feições espectrais (comportamento espectral) e de outros elementos auxiliares, como forma, tamanho, textura, etc.

Assim, a estrutura da paisagem, comumente utilizada como dado complementar aos estudos de interpretação de imagens, pode ser obtida por meio da ecologia da paisagem

utilizando métricas que auxiliam na compreensão da complexidade de uma paisagem e a forma como esta influencia determinadas relações ecológicas (CARRÃO et al., 2001).

As métricas associadas às ferramentas computacionais, para extrair informações de imagens, como mineração de dados espaciais e espectrais podem ser utilizadas para aumentar o potencial da análise e aplicações de dados de sensoriamento remoto (KORTING et al., 2008) tornando-se ferramentas importantes para tomadas de decisões.

No presente estudo utilizaram-se técnicas de mineração de dados e métricas de ecologia da paisagem para classificar e quantificar diferentes tipos de cobertura florestal presentes no município de São Luis do Paraitinga, São Paulo no ano de 2011.

2. Metodologia de Trabalho

Esta pesquisa foi realizada no município de São Luís do Paraitinga (Figura 1), localizado no estado de São Paulo, no Médio Vale do Paraíba do Sul, a uma latitude de 23°13'18" Sul e uma longitude 45°18'36" Oeste.

O município ocupa uma área de aproximadamente 621,4 Km², apresenta uma altitude média de 741 metros caracterizando-se geomorfologicamente por ser uma depressão de origem tectônica com declividade variável. Esta região é considerada de transição entre a floresta ombrófila densa (da encosta) e estacional semidecidual (do interior), com alguma influência da floresta ombrófila mista (CEIVAP, 2011; IBGE, 2011; CANTINHO; DIAS; BATISTA, 2011).

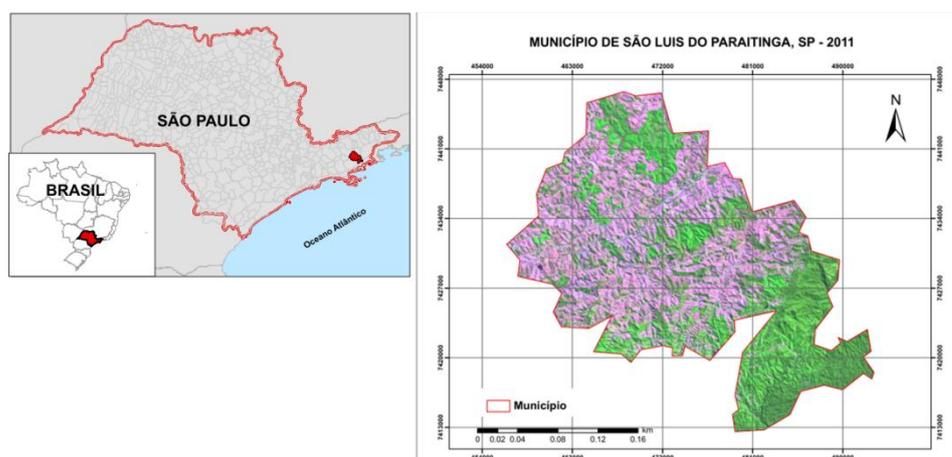


Figura 1. Representação da área de estudo por meio de composição colorida das bandas 5 (R) 4 (G) 3 (B) da imagem TM/Landsat-5 de 2011.

Para realizar essa pesquisa foram utilizadas imagens TM/Landsat-5 (resolução espacial de 30 metros) referentes à órbita/ponto 218/76, de 9 de setembro de 2001 e de 5 de setembro de 2011. O registro dessas cenas foi realizado com a imagem de referência disponível na base de dados *Global Land Cover Facility: Earth Science Data Interface – GLFC* (<<http://glcfapp.glc.umd.edu:8080/esdi/index.jsp>>) de mesma órbita/ponto. Essa etapa foi realizada no aplicativo computacional ENVI 4.7, com ajuste polinomial de 1º grau e erro menor que 0,5 pixel, ou seja, menor de 15 m, o que é aceitável para esse tipo de análise. Todo o processo foi realizado em sistema de projeção UTM, zona 23 Sul e Datum WGS-84.

Imagens *GeoEye* de alta resolução espacial visualizadas por meio do aplicativo *Google Earth 6.1.0.4857*, foram utilizadas como mapa base para auxílio na interpretação dos alvos nas imagens TM/Landsat-5.

Inicialmente, as imagens originais de 30 metros passaram por processo de correção atmosférica, por meio da correção radiométrica, utilizando a conversão das imagens em Número Digital para Reflectância no topo da atmosfera e posteriormente para reflectância de

superfície, em que utilizou a técnica do pixel escuro. Ambas as operações foram realizadas no aplicativo computacional ENVI 4.7.

Nessa etapa também foi utilizado a imagem NDVI (Normalized Difference Vegetation Index), com a finalidade de obter a máscara da cobertura florestal no município em estudo. Os valores do NDVI foram obtidos a partir da Equação 1 (Rouse et al., 1974)

$$NDVI = \frac{(NIR-RED)}{(NIR+RED)} \quad (1)$$

em que, NIR e RED representam os valores de radiância para as bandas do infravermelho próximo e vermelho, respectivamente.

Assim como Gonçalves et al. (2006), após gerar as imagens NDVI com valores entre -1 e 1, aplicou-se um fatiamento para realçar apenas as áreas com vegetação. Para obter as áreas de cobertura florestal agrupou, por meio do fatiamento, os pixels da imagem cujos valores do NDVI situavam entre 0,55 a 1,0.

Finalizando essa etapa, a máscara com a cobertura florestal do município recebeu um filtro de mediana 5 x 5 para retirar pixels isolados através da ferramenta (convolution) e o resultado foi transformado em vetores no formato de polígonos. Todo o processamento do NDVI foi realizado na imagem bruta (30 m) dentro do aplicativo ENVI 4.7.

Para facilitar a interpretação das classes nas imagens TM/Landsat-5, fez-se a restauração das mesmas (FONSECA, 1988) e foram reamostradas para pixels de 10 m, conforme Moreira (2007). Este procedimento foi realizado no aplicativo SPRING 4.3.3. (CÂMARA et al., 1996). Sobre as imagens restauradas aplicaram-se as máscaras de vegetação como demonstrado na Figura 2.

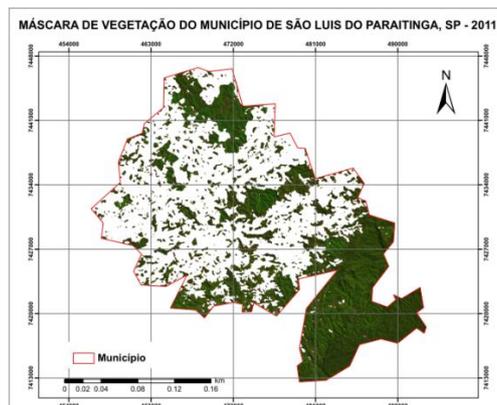


Figura 2. Máscara de vegetação florestal no ano de 2011.

A segunda parte do trabalho foi realizada dentro do aplicativo computacional GeoDMA (Geographical Data Mining Analyst) (KORTING et al., 2008) para realizar uma análise orientada a objeto. Este sistema manipula e visualiza dados armazenados em bancos de dados geográficos (KORTING et al., 2008) e permite extrair dados espaciais e espectrais que são calculados e armazenados no aplicativo computacional TerraView.

Para o presente estudo foram utilizadas 12 (doze) métricas entre espaciais e espectrais no GeoDMA, são elas: número de polígonos (NP), área (AREA), perímetro (PERIM), relação perímetro-área (PARA), circularidade (CIRCLE), índice de forma (SHAPE), retangularidade (RECT), elipticidade (ELLIPTIC), densidade, moda (MODE), média (MEAN) e desvio padrão (STDDEV).

Foram definidas quatro classes de cobertura vegetal, sendo elas: Eucalipto, Fragmento Florestal, Mata Ciliar e Outros (Figura 3). A classe Outros foi definida para aqueles alvos de difícil discriminação e com alta confusão na interpretação.

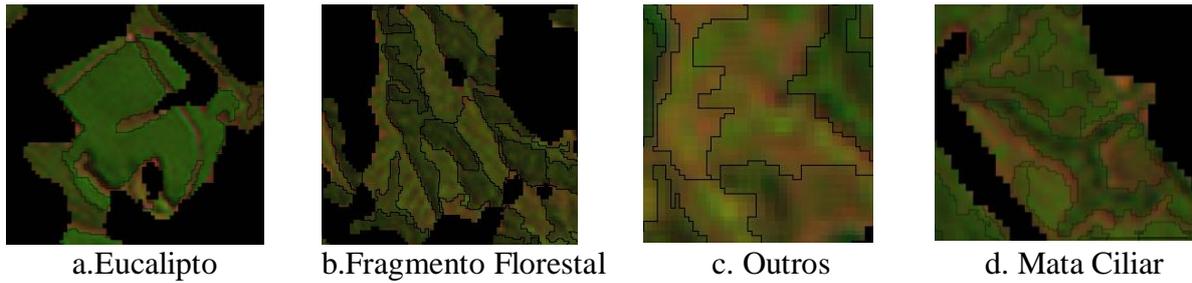


Figura 3. Exemplos de amostras de treinamento.

Aplicou-se o algoritmo de segmentação baseado na distância euclidiana e área mínima. Para isso foram testadas diversas segmentações, entretanto o limiar que melhor se enquadrou nos dados em análise foram os valores de 9 e 300 para distância euclidiana e área mínima, respectivamente. Cada objeto gerado possui como atributo os valores calculados a partir das métricas selecionadas. O próximo passo foi selecionar amostras de treinamento (Figura 3), que resultou num conjunto de informações. Estas amostras foram utilizadas para a classificação baseada em árvores de decisão pelo algoritmo C4.5.

Para validar a classificação foram coletadas amostras de referência com auxílio das imagens *GeoEye* de alta resolução espacial. Essas amostras foram coletadas de forma estratificada regular (Figura 4) onde cada vértice de uma malha de 2 x 2 km representou um ponto amostral, totalizaram 94 amostras.

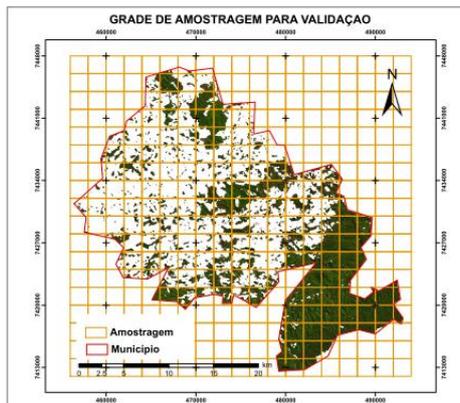


Figura 4. Grade de amostragem estratificada regular (alinhada) de 2 x 2 km.

Com as amostras de referência foi realizado, por meio da matriz de confusão, o cálculo do coeficiente de concordância Kappa (Equação 2), da Exatidão Global ou Desempenho Geral (Equação 3), da Acurácia do Produtor (Equação 4) e da Acurácia do Usuário (Equação 5).

$$K = \frac{D-Q}{D-T} \tag{2}$$

em que, D é o elemento da diagonal principal e T é o número total de amostras.

$$Q = \sum_{j=1}^r X_{i+} X_{+i}$$

em que, X_{i+} representa o total da linha para uma dada categoria informacional e X_{+i} representa o total da coluna para uma dada categoria informacional.

$$DG = \frac{D}{T} \tag{3}$$

Para os cálculos dos índices de acurácia foram utilizadas as seguintes equações:

$$\text{Acurácia do produtor} = \frac{X_{ij}}{X_{i+}} \quad (4)$$

$$\text{Acurácia do usuário} = \frac{X_{ij}}{X_{+j}} \quad (5)$$

em que, X_{ij} representa o elemento de determinada classe na linha i e coluna j .

3. Resultados e discussão

Após a realização das etapas iniciais deste trabalho, o classificador C4.5 gerou a partir das métricas calculadas as árvores de decisão utilizando as métricas que melhor se enquadraram para cada conjunto de dados (Figura 5).

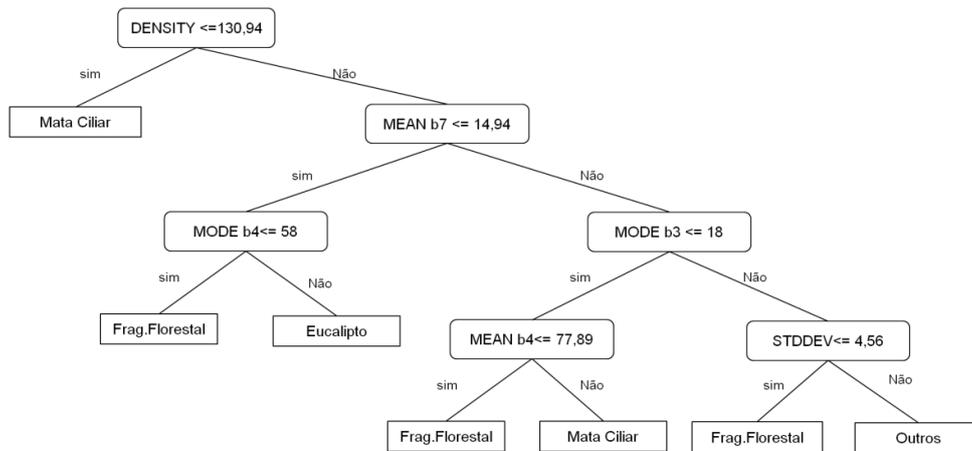


Figura 5. Árvores de decisão.

Utilizando alguns dos atributos fornecidos pela árvore de decisão foi gerado o gráfico de dispersão das classes (Figura 6) para melhor compreender como ocorre a separação entre elas. Observou-se a utilização das métricas espectrais: média da banda 7 e moda da banda 4, sendo possível visualizar apenas as classes Eucalipto e Fragmento Florestal, confirmando a necessidade de outros parâmetros para individualizar todas as classes.

Utilizando a árvore de decisão gerada pelo algoritmo de classificação obteve-se o mapa de cobertura florestal, apresentado na Figura 7, com base na máscara de vegetação inserida pelas etapas anteriores na pesquisa. Com as Figuras 7 e 8 observa-se que a cobertura florestal refere-se a 8,6% com Eucalipto e, 36,1% com Fragmentos Florestais.

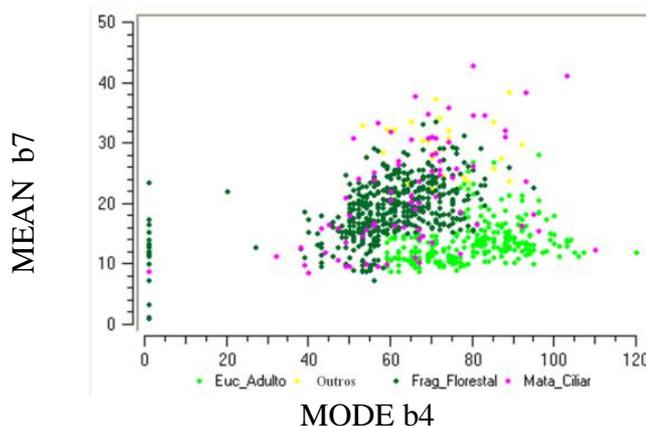


Figura 6. Gráfico de dispersão utilizando atributos espectrais.

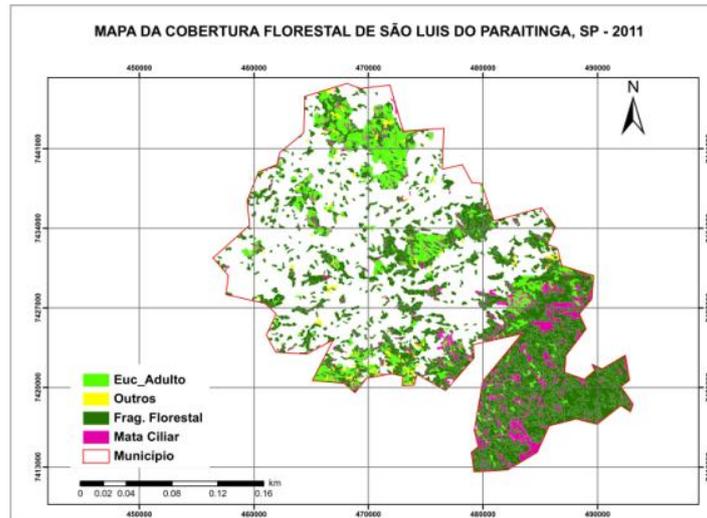


Figura 7. Mapas de cobertura florestal de São Luis do Paraitinga, SP.

Os resultados encontrados nessa pesquisa indicam que a importância do setor florestal, com grande participação no número de fragmentos. Esse tipo de análise é importante para dar continuidade aos estudos realizados por Arguello (2010) e Cantinho; Salgado; Batista (2011) que indicaram o incremento de eucalipto nesse município nos períodos entre 2000 a 2007 e 2007 a 2010, respectivamente.

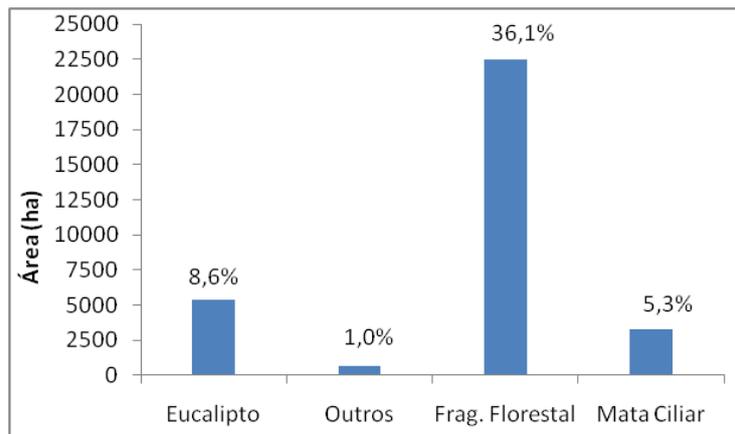


Figura 8. Área (ha) da cobertura florestal no município de São Luis do Paraitinga, SP.

Análises dessa qualidade são importantes para a gestão pública, bem como no monitoramento, aplicação e fiscalização da ação pública implementada pelo instrumento nº 759.170-5/3-00 TJSP, de 28 de março de 2008, cujo objetivo era conter os impactos e os problemas ocasionados pela expansão dos cultivos de Eucalipto no município.

As métricas CIRCLE, PARA E SHAPE são importantes para compreender a estrutura dos polígonos e como esses se relacionam com o ambiente. Em geral, as classes apresentaram forma mais irregulares com valores de SHAPE afastando-se de 1 (valores iguais a 1 indicam forma mais compacta). Contudo estas se apresentaram com pouca complexidade de forma com valores de PARA próximo de zero. Essas métricas são confirmadas também pela CIRCLE, cujos valores próximos a 1 afastam a forma de um círculo (Tabela 1).

Vale ressaltar que a grande porcentagem de fragmentos florestais da Mata Atlântica encontrada nesse município confirmou o exposto por Ribeiro et al. (2009), cujas áreas desse bioma que ainda estão preservados são decorrentes da dificuldade de acesso como no caso da Serra do Mar, localizada a Sudeste da área de estudo.

Tabela 1. Médias das métricas de paisagem calculadas

Classes\métricas	NP	AREA (ha)	CIRCLE	PERIM (km)	PARA	SHAPE
Total 2011						
Eucalipto Adulto	486	10,45	0,75	2,64	0.03	2.11
Outros	80	6,09	0,58	1,93	0.03	1.70
Frag. Floresta	1971	10,69	0,74	2,70	0.03	2.09
Mata Ciliar	354	6,71	0,91	2,72	0.05	2.76

A validação da classificação realizada pelo algoritmo C4.5 por meio do índice Kappa resultou em valores de 0.80, a partir da matriz de confusão que pode ser observada pela Tabela 2. Apesar de classificações sobre esse índice estarem reduzidas pode-se afirmar que segundo Cogalton e Green (1998) esses valores indicaram uma classificação razoável. Dessa forma, é possível identificar que o classificador realizou poucas confusões entre as classes, oferecendo confiabilidade e aceitação para a classificação realizada.

Tabela 2. Matriz de confusão das amostras de treinamento

		Classificação 2011					
		Classes	Eucalipto	Outros	Frag. Florestal	Mata Ciliar	Total
Referência	Eucalipto		17	0	1	1	19
	Outros		2	3	0	0	5
	Frag. Florestal		1	0	60	3	64
	Mata Ciliar		0	0	1	5	6
	Total		20	3	62	9	94

Os bons resultados obtidos com as classificações pelo índice kappa são confirmados com a Exatidão Global de 90% e com os altos valores das acurácias do produtor e usuário (Tabela 3). Observa-se, por exemplo, 85% da classe eucalipto foram corretamente classificadas como eucalipto.

Tabela 3. Acurácia do produtor e do usuário

	Eucalipto	Outros	Frag. Florestal	Mata Ciliar
Acurácia Produtor (x_{ii}/x_{i+})	85%	100%	96.7%	55.5%
Acurácia Usuário (x_{ij}/x_{+j})	89.4%	60%	93.7%	83.3%

4. Conclusões

A geração de classificações pelo método de mineração de dados associada a métricas de ecologia da paisagem mostrou-se uma ferramenta acessível e confiável para extrair atributos espaciais e espectrais de dados de sensoriamento remoto. Essa metodologia foi eficaz na separação de classes de coberturas florestais e na determinação de parâmetros importantes para reconhecer a diferença existente entre elas.

As métricas utilizadas foram importantes para compreender melhor os objetos inseridos na paisagem e podem servir de auxílio na identificação de pressões, como as alterações na cobertura florestal.

5. Referências Bibliográficas

Arguello, F. V. P. **Expansão do eucalipto no trecho paulista da bacia hidrográfica Paraíba do Sul**. 2010. 79 p. Dissertação (Mestrado em Ciências Ambientais) - Universidade de Taubaté, Programa de Pós-graduação em Ciências Ambientais, Taubaté. 2010.

Calegari, L.; Martins, S. V.; Gleriani, J. M.; Silva, E.; Busato, L. C. Análise da dinâmica de fragmentos florestais no município de Carandaí, MG, para fins de restauração florestal. **Revista Árvore**, Viçosa-MG, v.34, n.5, p.871-880, 2010.

Câmara, G.; Souza, R. C. M.; Freitas, U. M.; Garrido, J. **SPRING: Integrating remote sensing and GIS by object-oriented data modeling. Computers & Graphics**, v.20, n.3, p.395-403, 1996.

Cantinho, R. Z.; Salgado, M. P. G.; Batista, G. T. Análise da expansão do eucalipto no município de São Luís do Paraitinga, SP, Brasil. In: Simpósio Brasileiro de Sensoriamento Remoto, 15. (SBSR), 2011, Curitiba. **Anais...** São José dos Campos: INPE, 2011. p. 2083-2090. DVD, Internet. ISBN 978-85-17-00056-0 (Internet), 978-85-17-00057-7 (DVD).

Carrão, H.; Caetano, M.; Neves, N. **LANDIC: cálculo de indicadores de paisagem em ambiente SIG**. In: Encontro de Utilizadores de Informação Geográfica - ESIG 2001, 6., Oeiras, Portugal, 28-30 nov., 2001. Disponível em: <<http://esig2001.tripod.com>>. Acesso em: 18 nov. 2011.

CEIVAP. Comitê de Integração da Bacia Hidrográfica do Rio Paraíba do Sul. Dados geoambientais. Disponível em: <http://ceivap.org.br/bacia_1_2.php>. Acesso em: 08 out. 2011.

Congalton, R. G.; Green, K.; **Assessing the accuracy of remotely sensed data: principles and practices**. New York: Lewis Publishers, 1998. 137p.

Freitas, S. R.; Hawbaker, T. J.; Metzger, J. P. Effects of roads, topography, and land use on forest cover dynamics in the Brazilian Atlantic Forest. **Forest Ecology and Management** 259 410–417, 2010.

Gonçalves, D. A.; Júnior, W. C. S.; Morelli, F. Identificação e delimitação de áreas queimadas no Pantanal. In: SIMPÓSIO DE GEOTECNOLOGIAS NO PANTANAL, 1. (GEOPANTANAL), 2006, Campo Grande. **Anais...** Campinas: Embrapa Informática Agropecuária; São José dos Campos: INPE, 2006. p. 622-627. CD-ROM. ISBN 85-17-00029-3. Disponível em: <<http://urlib.net/sid.inpe.br/mtc-m17@80/2006/12.12.12.38>>. Acesso em: 18 nov. 2011.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Banco de Dados Agregados**. Disponível em: <http://www.ibge.gov.br>. Acesso em: 08 out. 2011.

Korting, T. S.; Fonseca, L. M.; Escada, M. I. S.; Silva, F. C.; Silva, M. P. S. **GeoDMA: a novel system for spatial data mining**. IEEE International Conference on Data Mining Workshops, 2008.

Moreira, M. A. **Fundamentos do sensoriamento remoto e metodologias de aplicação**. 4 ed, Viçosa-MG. Ed. UFV, 2011, 422p.

Ribeiro, M. C.; Metzger, J. P.; Martensen, A. C.; Ponzoni, F. J.; Hirota, M. **M. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation**. Biological Conservation, v.142, p.1141–1153, 2009.

Rouse, J. W.; Haas, R. H.; Schell, J. A.; Deering, D. W.; Harlan, J. C. **Monitoring the vernal advancement of retrogradation (greenwave effect) of natural vegetation**. NASA/GSFC, Type III, Final Report, Greenbelt, MD, 1974, 371 p.

Saito, E. A.; Korting, T. S.; Fonseca, L. M. G.; Escada, M. I. S.; Mineração em dados espaciais de desmatamento do PRODES utilizando métricas da paisagem caso de estudo município de Novo Progresso- PA. In: Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, 3., 2010. Recife. **Anais...** 2010, p. 001-009.