CLASSIFICAÇÃO DO ESTÁDIO SUCESSIONAL DA VEGETAÇÃO EMPREGANDO MINERAÇÃO DE DADOS E SOFTWARES OPEN SOURCE

Classification of sucessional forest stages using data mining and open source softwares

Resumo:

Este trabalho teve como objetivo avaliar diferentes aplicativos *open source* relacionados à análise baseada em objeto para a classificação de estádios sucessionais de florestas secundárias da Floresta Ombrófila Mista (FOM) em duas áreas-teste no planalto sul catarinense. Foram utilizadas ortoimagens do Sistema Aerotransportado de Aquisição e Pós-processamento de Imagens (SAAPI) de alta resolução espacial (0,39 m). Os dados consistem de três bandas no espectro visível (0,38 - 0,70 μm), três bandas no espectro do infravermelho próximo (0,76 - 0,78 μm) e o Modelo Digital de Superfície (MDS). As metodologias foram desenvolvidas utilizando mineração de dados com algoritmos de árvore de decisão e seleção de atributos nos aplicativos InterIMAGE, WEKA e QGIS. Os resultados se mostraram satisfatórios para classificar estádios sucessionais da FOM, assim como outras classes de uso e cobertura da terra. As classificações apresentaram um índice Kappa variando entre 0,53 e 0,59 e Kappa condicional variando entre 0,29 e 0,83 para os estádios sucessionais da vegetação. Estes resultados demonstram o potencial dessas abordagens na extração de informações de imagens de alta resolução espacial, bem como a possibilidade de fornecer subsídios para a implementação de políticas públicas e no monitoramento dos recursos florestais.

Palavras-chave: Sucessão florestal; Árvores de decisão; Imagens Aerotransportadas, Atributos.

Abstract:

This study aimed to evaluate open source softwares in order to classify secondary successional forest stages in Ombrophilous Mixed Forest (FOM) environments in Southern Brazil. Two test sites were selected in the mountainous region of Santa Catarina State (SC). We used scenes from the airborne system for acquisition and post-processing of images (SAAPI) with a spatial resolution of 0.39m. The dataset consists of orthorectified images containing three spectral bands in the visible range (i.e. 0.38–0.70μm), three spectral bands in the near infrared (i.e. 0.76–0.78μm) and a digital surface model. The methodologies were developed using feature selection and decision tree algorithms in the following open source softwares: InterIMAGE, WEKA and QGIS. The results were satisfactory to classify successional stages of FOM as well as other classes of land use and land cover. The obtained Kappa indices ranged from 0.53 to 0.59 and the conditional Kappa varied from 0.29 to 0.83 for the successional forest stages. These results demonstrate the potential of these approaches for the extraction of information in high spatial resolution imagery as well as the possibility of providing subsidies for the implementation of public policies and monitoring of forest resources.

Keywords: Secondary forests; Decision tree; Aerophotogrammetric survey; Features.

1. Introdução

O mapeamento dos remanescentes florestais da Mata Atlântica e seus estádios de sucessão consiste em uma etapa fundamental para implantação de diversos estudos, ações de fiscalização e manejo ambiental (Ribeiro et al., 2009). A proposição de metodologias que contemplem essa temática permite avaliar quantitativamente e qualitativamente os fragmentos remanescentes bem como a sua distribuição espacial. Essas informações podem subsidiar estudos complementares, a exemplo do estabelecimento de áreas prioritárias para conservação, entre outros (Amaral et al., 2009). Uma vez que os estádios sucessionais são mencionados nos textos das leis de proteção ambiental, a exemplo da Lei da Mata Atlântica (Lei 11.428/2006) (Brasil, 2006), torna-se, portanto, necessário avaliar o papel de dados remotamente situados para atendimento da legislação.

No mapeamento de estádios de sucessão florestal, os espectros de reflectância são muito similares, tornando o processo de classificação uma tarefa desafiadora (Vieira et al., 2003). Portanto, a pesquisa de técnicas e metodologias de classificação que contemplem atributos relacionados à forma do alvo, à textura ou relações entre alvos vizinhos no mapeamento das fases de sucessão florestal ainda é necessária.

Neste contexto, o advento de novos sistemas sensores e a melhoria das resoluções radiométrica, espectral, temporal e espacial impulsionaram o surgimento de novas metodologias para extração de informações, superando as limitações das metodologias tradicionais existentes (*pixel-a-pixel* e por regiões). A segmentação das imagens ganhou destaque quando foi incorporada à análise baseada em objeto (*Object-based Image Analysis*- OBIA) (Blaschke e Strobl, 2001). Nela, os *pixels* não só são agrupados em segmentos, mas também são reconhecidos como objetos. Os objetos, diferentemente das regiões ou segmentos, são dotados de significado e identidade, sendo distinguíveis, portanto, pela sua própria existência e não pelas propriedades que possuem. Além das propriedades espectrais, espaciais e texturais, os objetos apresentam relações contextuais e semânticas, que podem ser utilizadas para a análise de imagens e aproximam-se dos processos cognitivos humanos de interpretação de imagens.

A rede semântica, que expressa um modelo de conhecimento e se materializa graficamente em uma estrutura hierárquica de classes, pode ser gerada de forma heurística pelo usuário, testando iterativa e interativamente os descritores, as funções e os seus limiares para a discriminação adequada dos alvos. Pode também ser gerada de forma automática, através da aplicação de técnicas de mineração de dados (Francisco e Almeida, 2012). Estas consistem na extração de conhecimento de uma base com um grande volume de dados por meio de métodos inteligentes. O modelo derivado da mineração pode ser representado de várias formas, entre elas, as árvores de decisão, representadas por um fluxograma com estrutura de árvore e facilmente convertidas em regras de classificação (Han e Kamber, 2006).

Este estudo visou à classificação do uso e cobertura da terra em ortoimagens produzidas pelo Aerolevantamento Fotogramétrico de SC e, através de técnicas de mineração de dados, identificar os principais atributos que diferenciam os estádios sucessionais da vegetação em áreas de Floresta Ombrófila Mista. Para isto, testaram-se algoritmos de árvore de decisão e seleção de atributos, nos sistemas *open source Interpreting Images Freely* (InterIMAGE), *Waikato Environment Knownledge Analysis* (WEKA) e *Quantum Geographic Information System*

(QGIS). Com isso, pretendeu-se analisar a aplicabilidade das informações extraídas para as áreas de fiscalização, gestão, manejo e recuperação ambiental.

2. Material e Métodos

O estudo foi realizado em duas áreas pertencentes à região fitoecológica de Floresta Ombrófila Mista (Klein, 1978), situada na mesorregião Serrana e microrregião Campos de Lages do estado de Santa Catarina (SC) (IBGE, 2015). A área-teste A está localizada no município de Anita Garibaldi nas coordenadas geodésicas médias 27º41'21" sul e 51º07'48" oeste. A área-teste B localiza-se em Urubici, no interior do Parque Nacional de São Joaquim, nas coordenadas geodésicas médias 28º00'54" sul e 49º35'30" oeste. Segundo a classificação de Köppen, o clima das duas áreas-teste é do tipo "Cfb".

O aerolevantamento fotogramétrico no estado de Santa Catarina foi executado com recursos da Secretaria de Estado do Desenvolvimento Econômico Sustentável (SDS) entre 2010 e 2011, e se produziram aproximadamente 57 mil ortoimagens. Esse levantamento foi realizado com um Sistema Aerotransportado de Aquisição e Pós-processamento de Imagens Digitais (SAAPI), com sensor CCD (*Charge Coupled Device*, "dispositivo de carga acoplado"), resolução geométrica de 39 cm, e filtro UV-Sky, que filtra a luz ultravioleta e compensa o efeito de névoa atmosférica (Piazza, 2014).

Os produtos gerados pelo aerolevantamento são a composição em cores verdadeiras nos canais do vermelho, verde e azul (*Red, Green, Blue* - RGB) e a composição colorida utilizando o infravermelho próximo (*Near Infrared* - NIR). O levantamento também obteve o modelo digital de terreno (MDT) e o modelo digital de superfície (MDS). Os dados foram recebidos pela SDS com as etapas de pré-processamento, ajuste radiométrico, níveis de contraste, tonalidade, homogeneização das imagens, balanceamento de cores e ortorretificação já realizadas (Engemap, 2012).

Os dados de entrada utilizados no presente estudo foram as ortoimagens correspondentes às bandas espectrais do visível e infravermelho próximo e o MDS, formando um conjunto contendo sete bandas. As ortoimagens foram recortadas em 1.500x1.500 *pixels*, visando maior agilidade e redução do custo computacional na aplicação das metodologias propostas. Além disso, alguns dos aplicativos utilizados, como o *Segmentation Parameters Tuner* (SPT) e InterIMAGE, apresentam limitações ao trabalharem com imagens de grandes dimensões espaciais nos casos de processamento em PCs convencionais, que não dispõem de placa gráfica para processamento e/ou de grande número de processadores paralelos e quantidade excedente de memória volátil (RAM).

Ponzoni et al. (2012) falam da necessidade da conversão dos números digitais (ND) das imagens para valores de reflectância de superfície, para possibilitar a caracterização espectral dos alvos, já que um valor de ND de uma imagem em uma banda específica não está na mesma escala de outro ND de outra imagem ou outra banda espectral. Não foi possível adquirir informações detalhadas em relação à calibração do sensor ou do aerolevantamento.

Foi elaborado um mapa correspondente à verdade terrestre de cada área-teste. Para isso, um fotointérprete fez a interpretação das ortoimagens utilizando o MDS e a ferramenta *3D Analyst* do ArcGIS. Entre as principais classes identificadas visualmente nas imagens, estão: vegetação em estádio inicial (VEI), médio (VEM) e avançado (VEA), reflorestamento com espécie exótica *Pinnus spp*., cultivos agrícolas (agricultura), campo, campo sujo (campo com algum tipo de vegetação esparsa) e sombra.

A caracterização da vegetação também considerou dados de campo obtidos no Inventário Florístico Florestal de SC (Vibrans et al., 2012) para a área-teste A, e dados de Faxina (2014) para a área-teste B. Nestes levantamentos, o estádio da vegetação foi definido conforme os critérios da Resolução CONAMA nº 04/94 (Brasil, 1994), que considera: diâmetro a altura do peito (DAP); altura das árvores; área basal; estratos predominantes; espécies indicadoras; diversidade e dominância de espécies; cobertura do dossel; presença e características da serapilheira e sub-bosque; existência, diversidade e quantidade de epífitas e trepadeiras.

O algoritmo segmentador proposto por Baatz e Schäpe (2000) foi escolhido pela velocidade de execução e capacidade de extração de objetos homogêneos em uma mesma escala. Nele, são estabelecidos parâmetros de cor, forma e escala. De acordo com Gao et al. (2011), o tamanho médio dos objetos na imagem tem um impacto significativo na acurácia da classificação. Para a escolha dos parâmetros da segmentação, foram avaliados nove cenários, compostos pela combinação de três fatores de escala (60, 80 e 100), com pesos de cor e forma variando de 0,3 a 0,7.

A avaliação dos resultados da segmentação foi feita no aplicativo SPT (Achanccaray Diaz, 2014), através de uma função de aptidão, ou métrica, que indica a qualidade da segmentação em função dos segmentos de referência criados pelo usuário. Neste trabalho, optou-se pelo *Reference Bounded Segments Booster* (Assistente para Segmentos Delimitados como Referência - RBSB), métrica proposta por Feitosa et al. (2006), utilizada por Feitosa et al. (2009), Novack (2009) e Leonardi (2013), que corresponde à razão entre a área de dois segmentos fora da interseção com a área de referência. Quanto mais próximo de zero, mais a segmentação gerada se aproxima da segmentação de referência, sendo que zero corresponde ao ajuste perfeito. A métrica RBSB mostra boa correlação com a percepção humana de qualidade de segmentação (Feitosa et al., 2006).

Para uma das classificações, usou-se o InterIMAGE, que é um *software* de domínio público e código aberto (Costa et al., 2008). Ele dispõe do algoritmo C4.5, que possibilita que a classificação baseada em objeto seja automatizada, e não somente através da rede semântica. Diferentemente do *software* WEKA, o InterIMAGE não oferece ao analista a possibilidade de manipular o tamanho da árvore de decisão, gerando árvores de diversas dimensões (Rodrigues, 2014).

O primeiro procedimento realizado no InterIMAGE foi a construção da rede semântica, que representa as classes que se espera encontrar na cena. Neste trabalho, criaram-se redes operacionais, ou seja, sem relação hierárquica entre as classes, já que o objetivo foi explorar a classificação automatizada. Desta forma, cada classe (nó-folha) foi associada ao mesmo nó-pai, sem níveis intermediários.

Na ferramenta *Samples editor* (editor de amostras), foi feita a segmentação das imagens com o algoritmo Baatz e Schäpe, e procedeu-se à coleta aleatória das amostras. As amostras foram coletadas de forma a abranger qualquer variação interna das classes quanto à cor, tonalidade, forma, textura e brilho, e também considerando a representatividade da classe nas imagens.

Em seguida, estipularam-se os atributos a serem extraídos de cada segmento para serem usados na classificação. Foram gerados 47 atributos, sendo 43 deles espectrais e quatro operações entre bandas espectrais, estas últimas escolhidas por explorar o contraste que a vegetação apresenta entre as bandas do visível e do infravermelho próximo. Maiores detalhes sobre os atributos utilizados podem ser encontrados em Autor (2015). Foram priorizados atributos estatísticos ao invés dos espaciais, por se tratar predominantemente de áreas naturais, em que os objetos têm formas irregulares. Como citado por Yu et al. (2006), diferentemente da classificação de áreas urbanas, características geométricas têm pouca contribuição para a classificação da vegetação em

imagens de alta resolução espacial, já que esta não possui um padrão espacial óbvio que poderia ser evidenciado na classificação.

A classificação foi feita com o algoritmo *top down TA_C45_Classifier*, que utiliza o conceito de árvore de decisão proposto por Quinlan (1993). O operador *TA_C45_Classifier* foi associado a apenas um dos nós-filho, correspondente à classe VEA da rede semântica de cada área-teste. Neste nó, habilitou-se a opção *Multiclass*, ficando responsável por repassar as hipóteses aos demais nós da rede. A todos os nós restantes, atribuiu-se o operador *Dummy top-down*. Neste operador, nenhuma hipótese é criada, e as informações são apenas repassadas de nó-pai para nó-filho (Rodrigues, 2014). A classificação final resulta em um arquivo *shape*, e a árvore de decisão é gerada em um arquivo de texto *.txt*.

Adicionalmente, usou-se a ferramenta WEKA, que incorpora um conjunto de algoritmos de aprendizado de máquina que possibilita a extração do conhecimento. A metodologia desenvolvida no WEKA compreendeu duas etapas: seleção de atributos e geração de modelo de classificação por árvore de decisão. O banco de dados utilizado nesta tarefa foram as amostras e respectivos atributos gerados no InterIMAGE convertido para o formato *Attribute-Relation File Format* (Formato de Arquivo Atributo-Relação- ARFF).

Após o procedimento de conversão, procedeu-se à seleção de atributos envolvendo duas ferramentas a serem escolhidas pelo usuário: o avaliador de atributos e o método de busca. O avaliador determina qual método é usado para atribuir um valor a cada subconjunto de atributos, e o método de busca determina o tipo de busca a ser realizada. Utilizou-se o avaliador *Correlation-based Feature Selection* (Seleção de Atributos baseada em Correlação- CFS) associado ao método de busca *Best-First* (método da melhor busca inicial). O CFS considera um conjunto de atributos "bom", quando contém atributos altamente correlacionados com a classe e não-correlacionados entre si. A base deste método é uma heurística de avaliação de subconjuntos que considera não somente a utilidade de atributos individuais, mas também o nível de correlação entre eles (Karegowda et al., 2010).

A etapa seguinte à seleção de atributos foi a classificação supervisionada do banco de dados. Optou-se pela Árvore de Regressão e Classificação (*Classification and Regression Trees*- CART), proposta por Breiman et al. (1984), denominada *SimpleCart* no WEKA. Dentre os algoritmos de árvore de decisão, este foi o que gerou árvores de melhor acurácia e menor dimensionalidade.

Para proceder com a construção do modelo WEKA no QGIS, primeiramente converteu-se a árvore de decisão em regras, separando-as conforme a classe de uso e cobertura da terra à qual pertenciam. Utilizou-se no QGIS o arquivo *shape* resultante da segmentação e a extração de atributos, ambos gerados no InterIMAGE, correspondente a cada área-teste. Na tabela de atributos desse arquivo, selecionaram-se as feições que satisfaziam às regras correspondentes a cada classe, e o campo "*class*" era preenchido conforme a classe. Esse processo repetiu-se, mudando-se as regras de decisão conforme a classe. Assim, ao final do processo, todos os segmentos estavam rotulados com as respectivas classes.

Por fim, para a avaliação dos resultados, foram geradas matrizes de confusão, a partir de amostras aleatórias geradas no QGIS sobre os mapas de referência. O número de amostras variou conforme a área ocupada pela classe em cada imagem, porém, buscou-se obedecer ao número mínimo de 50 amostras, definido por Congalton e Green (1999).

Na matriz, são expressos os erros de omissão, ou seja, amostras que não foram classificadas de acordo com as classes de referência, e os erros de comissão, correspondentes a amostras de referência classificadas erroneamente como pertencentes a outras classes. A partir das matrizes, foram calculados os seguintes índices: (a) exatidão global - relação entre o número de amostras

classificadas corretamente sobre o número total de amostras de referência; (b) exatidão do produtor – relativa aos erros de omissão, a qual representa a relação entre o número de amostras classificadas corretamente da classe k e o número total de amostras de referência da classe k, (c) exatidão do usuário - referente aos erros de comissão, a qual representa a relação entre o número de amostras classificadas corretamente da classe k e o número total de amostras classificadas da classe k; (d) Kappa - analisa todos os elementos da matriz de confusão e (e) Kappa condicional do usuário (Kcu)- analisa a acurácia de acordo com a classe (Congalton e Green, 1999).

Foi executado o teste *z* para testar a significância estatística da diferença entre as classificações resultantes de cada metodologia (Congalton e Green, 1999). Atribuiu-se um nível de significância de 5% ($\alpha = 0,05$), com valor crítico de 1,96, ou seja, assumiu-se que se o valor do teste *z* fosse maior que o valor crítico haveria diferença significante entre os mapeamentos.

3. Resultados e Discussão

A partir do processamento no SPT, foram escolhidos os parâmetros que obtiveram valores de RBSB mais próximos de 0, ou seja, que mais se aproximaram aos segmentos de referência. O fator de escala selecionado para ambas as áreas foi de 80. O peso, tanto da cor quanto da forma, foi de 0,5 para área-teste A, e na área B, foi de 0,7 e 0,3 para cor e forma, respectivamente.

Na etapa de seleção de atributos, verificou-se que, dentre os 47 atributos gerados no InterIMAGE, o algoritmo CFS do WEKA selecionou subconjuntos com 15 atributos na área-teste A e 18 para a área-teste B (Tabela 1).

Verifica-se na Tabela 1 que a subtração entre a banda espectral correspondente à região do infravermelho próximo (IR1) pela banda correspondente à região do vermelho (R), além do *Normalized Vegetation Difference Index* (Índice de Vegetação por Diferença Normalizada- NDVI), obtido pela expressão ((IR1-R)/(IR1+R)), foram atributos selecionados nas duas áreas-teste, mesmo tendo sido calculados em função dos ND. Apesar da impossibilidade de se realizar a conversão para valores de reflectância, percebe-se que os espectros de ND (Figura 1A) mostraram-se coerentes com os reportados em Vieira et al. (2003); e Piazza (2014).

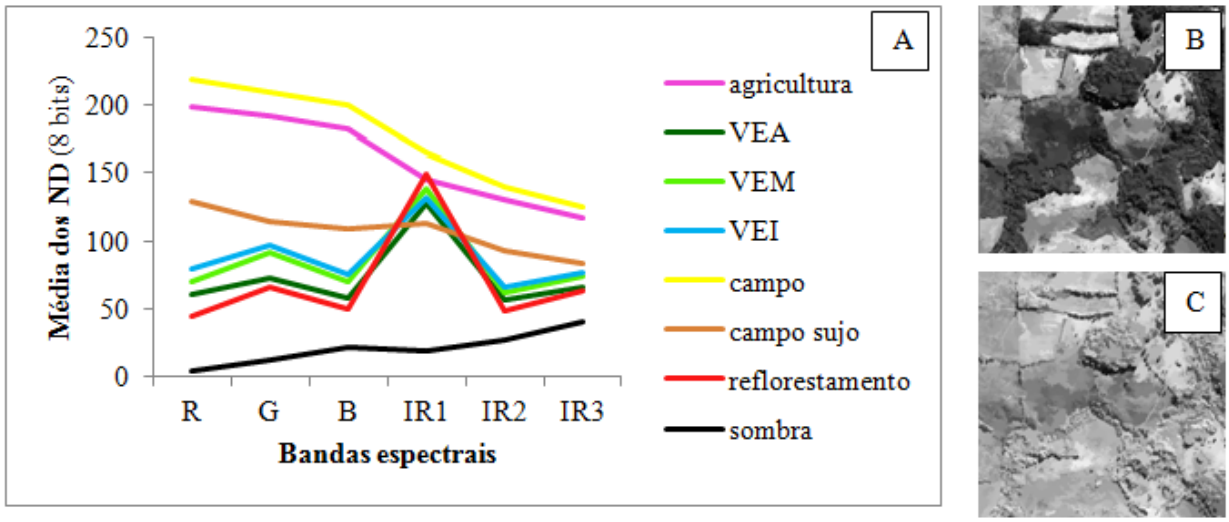

Figura 1: Perfil espectral das classes de uso e cobertura da terra da área-teste A (A) e espacialização da banda R (B) e banda IR1 (C) na ortoimagem da área-teste A.

Tabela 1: Atributos selecionados pelo algoritmo CFS no WEKA.

Área-teste A	Área-teste B
Subtração da banda IR1 pela R	Divisão da banda R pela IR1
Divisão da banda IR1 pela R	Subtração da banda IR1 pela R
Média da banda IR1	Divisão da banda IR1 pela R
Média da banda IR2	Entropia da banda B
Média do MDS	Entropia da banda IR1
Máx. valor de pixel no MDS	Média da banda B
Mín. valor de pixel na banda IR3	Média da banda IR1
Mín. valor de pixel no MDS	Média da banda IR2
NDVI	Máx. valor de pixel na banda B
Razão da banda IR1	Máx. valor de pixel no MDS
Razão da banda IR2	Mín. valor de pixel no MDS
Razão da banda IR3	NDVI
Desvio padrão da banda G	Razão da banda B
Desvio padrão da banda B	Razão da banda IR2
Desvio padrão da banda IR1	Razão da banda IR3
	Desvio padrão da banda G
	Desvio padrão da banda B
	Desvio padrão do MDS

Entre as classes de cobertura vegetal, a VEI apresentou maiores valores de ND ao longo do espectro, em função da maior regularidade e homogeneidade dos dosséis, com exceção da região do IR1, em que VEM apresentou valores maiores que a VEI. Tal comportamento pode ser atribuído ao fato de que a VEI na área-teste A apresenta semelhanças com algumas áreas de campo sujo, classe esta que teve valores menores de ND na banda IR1. Para a VEA, observa-se que a maior heterogeneidade dos dosséis, com maior presença de sombras, resultou em ND menores que os outros estádios.

De acordo com Ponzoni et al. (2012), quando o dossel florestal é dividido em dois ou três estratos verticais, espera-se que ele apresente tonalidade mais escura em função das sombras dos que as demais classes da vegetação nas bandas do visível. Igualmente, espera-se ainda uma maior atividade fotossintética e tonalidade clara na banda do infravermelho próximo em razão do espalhamento múltiplo da radiação eletromagnética por parte das folhas. A Figura 1B ilustra a espacialização das médias da banda espectral vermelho (R), em que a vegetação aparece em tons escuros, e IR1 (Figura 1C), em tons claros, indicando, portanto, valores mais altos de ND. Tal comportamento justifica o fato de a subtração entre essas duas bandas e o NDVI terem sido atributos selecionados nas duas áreas-teste.

Outros atributos selecionados para todas as áreas-teste foram o máximo e mínimo valor de *pixel* do MDS correspondente a cada amostra. Como o MDS representa apenas a variação altimétrica das feições em função das altitudes, as alturas dos objetos não são conhecidas, porém, a informação de textura atrelada a essa componente é alta, desempenhando um importante papel ao discriminar as classes. Yu et al. (2006) citam que as características topográficas foram mais importantes para fins de discriminação da vegetação do que as características espectrais. De forma semelhante, Francisco e Almeida (2012) demonstraram a importância da inclusão de outras informações, não somente espectrais, para a discriminação das classes de cobertura da terra. Nesse estudo, o atributo declividade foi capaz de discriminar duas classes (afloramento rochoso e vegetação herbácea rala) que apresentavam respostas espectrais e texturais semelhantes.

A Tabela 2 mostra as características das árvores de decisão geradas em cada um dos aplicativos. Percebe-se que, para as duas áreas, a árvore de decisão *SimpleCart* aliada à etapa de seleção de atributos resultou em árvores mais compactas e com menor número de atributos em comparação ao C4.5. Durante a análise da árvore de decisão, os dados mais importantes se encontram nos nós-folhas mais próximos ao nó-raiz. As operações entre as bandas do IR1 e R foram os grandes-nós, ou seja, os atributos principais das árvores geradas para as duas áreas, demonstrando a relevância de tais atributos ao discriminar a vegetação das demais classes.

Tabela 2: Características das árvores de decisão geradas em cada aplicativo.

Software	**Algoritmo**	**Característica da árvore**	**Área A**	**Área B**
InterIMAGE	C4.5	Nº ramificações	52	24
		Nº atributos	17	10
		Nº nós	27	13
		Atributo principal	(IR1-R)	NDVI
WEKA	SimpleCart	Nº ramificações	17	15
		Nº atributos	6	6
		Nº nós	10	8
		Atributo principal	(IR1-R)	(RdivIR1)

A Figura 2 mostra o mapa classificado da área-teste A, gerado com o algoritmo C4.5 no InterIMAGE. A Tabela 3A mostra que, para este algoritmo, a Exatidão Global foi de 67% e o índice Kappa de 0,59, considerado bom. Quanto aos estádios sucessionais da vegetação, os resultados podem ser considerados excelentes para as classes VEI e VEM, com Kappa condicional de 0,8, e muito bom para a VEA, com Kappa condicional de 0,6.

A Tabela 3B mostra a matriz de confusão resultante da classificação com a árvore de decisão *SimpleCart* gerada no WEKA. A classificação alcançou Exatidão Global de 63% e índice Kappa de 0,56, considerado bom. Apesar de o Kappa ser inferior ao obtido para a classificação no InterIMAGE, ambos os métodos não diferiram significantemente no teste z. A classe VEM apresentou o melhor resultado entre os estádios da vegetação, com Kappa condicional do usuário de 0,83. Nas duas metodologias, a classe sombra apresentou baixa exatidão e Kappa condicional do usuário, já que teve outras classes erroneamente atribuídas a ela: VEM, VEA e campo sujo, principalmente.

Outro erro de classificação recorrente na área-teste A foi observado para a classe reflorestamento (Refl.), classificada predominantemente como VEA. A classe VEA, nas duas metodologias, apresentou menor Kappa condicional do usuário que as demais classes de vegetação, tendo as classes VEM e reflorestamento atribuídas a ela. A VEI apresentou pouca confusão com os demais estádios de vegetação, mas foi erroneamente classificada como campo sujo no InterIMAGE, e também como agricultura (Agric.) e campo no WEKA. Tal fato deve-se à maior presença de solo na VEI. Foi verificado que a regra de decisão referente a esta classe encontra-se no mesmo ramo que as classes campo sujo e agricultura no *SimpleCart*, demonstrando a semelhança espectral destas classes.

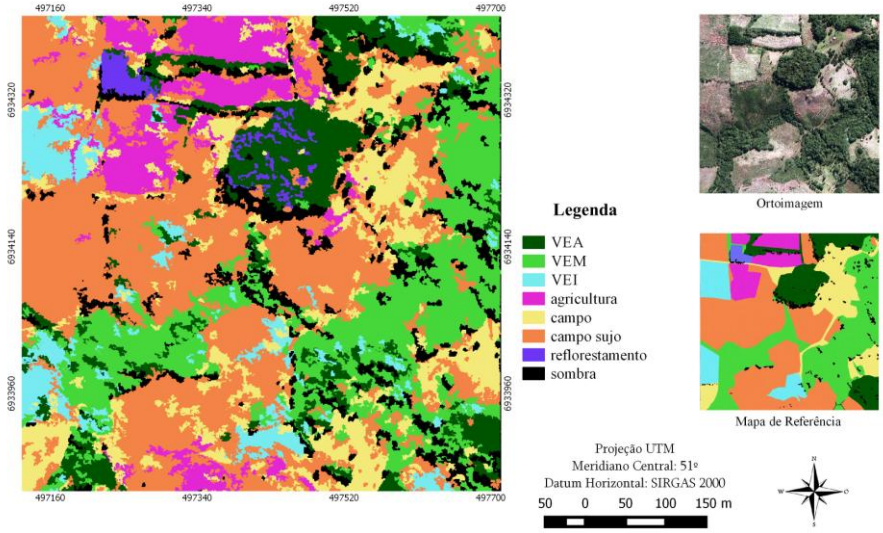

Figura 2: Classificação do uso e cobertura da terra da área-teste A com o algoritmo C4.5.

Tabela 3: Matriz de confusão da classificação da área-teste A no InterIMAGE com o algoritmo C4.5 (A); e no WEKA e QGIS com o algoritmo *SimpleCart* (B).

		Mapa de referência									
	Classe	VEI	VEM	VEA	Agric.	Campo	C. sujo	Refl.	Sombra	TCl	Kcu
Mapa classificado (A)	VEI	32	3	0	0	0	4	0	0	39	0,80
	VEM	0	79	6	0	2	5	0	1	93	0,81
	VEA	0	14	32	0	0	1	3	0	50	0,60
	Agric.	3	1	1	32	2	18	0	0	57	0,52
	Campo	2	6	1	2	44	20	0	0	75	0,53
	C. sujo	12	12	6	16	21	146	1	0	214	0,52
	Refl.	0	0	2	0	0	0	15	1	18	0,83
	Sombra	1	15	7	0	1	6	1	18	49	0,35
	TCo	50	130	55	50	70	200	20	20	595	0,59
	Ep	64%	61%	58%	64%	63%	73%	75%	90%	67%	
	Eu	82%	85%	64%	56%	59%	68%	83%	37%		
Mapa classificado (B)	VEI	30	5	2	0	2	13	1	0	53	0,53
	VEM	1	85	5	0	2	4	0	1	98	0,83
	VEA	1	12	34	0	0	5	8	1	61	0,51
	Agric.	5	1	2	43	6	37	0	0	94	0,41
	Campo	4	6	0	2	52	31	0	0	95	0,49
	C. sujo	9	6	4	5	7	103	0	0	134	0,65
	Refl.	0	0	1	0	0	0	10	0	11	0,91
	Sombra	0	15	7	0	1	7	1	18	49	0,35
	TCo	50	130	55	50	70	200	20	20	595	0,56
	Ep	60%	65%	62%	86%	74%	51%	50%	90%	63%	
	Eu	57%	87%	56%	46%	55%	77%	91%	37%		

*TCo= total de amostras coletadas; TCl: total de amostras classificadas; Ep= Exatidão do produtor; Ep= Exatidão do usuário; Kcu= Kappa condicional do usuário.

A classe campo sujo obteve bom Kappa condicional do usuário no WEKA, porém, apresentou altos erros de omissão, em vista de ter sido atribuída à classe campo, agricultura e VEI. O campo sujo apresenta uma variedade espectral intraclasse, já que possui áreas cobertas com uma vegetação esparsa, e outras mais "limpas", o que pode ter ocasionado o erro de classificação. Esse erro também pode estar associado ao mapa de referência desta classe, já que a área classificada como campo sujo no mapa de referência, em certos locais, pode tratar-se de áreas agrícolas abandonadas, assemelhando-se à classe agricultura. Quanto à confusão do campo sujo com a classe campo, percebe-se que ocorreram em regiões em que aquela classe possui maior brilho, semelhante ao observado por Novack (2009). Este autor, ao classificar áreas urbanas com imagens WorldView, obteve confusão entre as classes "solo exposto" e "vegetação rasteira". Ele atribuiu isto ao fato de as duas classes estarem misturadas em alguns locais, o que descaracteriza o comportamento espectral de ambas.

Interessante que a árvore de decisão *SimpleCart* do WEKA com apenas uma regra de decisão para cada uma das três classes de vegetação conseguiu índices de acerto considerados muito bons, equiparando-se ao C4.5, que utilizou quatro a cinco regras para cada estádio. Isto demonstra que árvores de decisão generalizáveis podem trabalhar melhor com a variabilidade dos dados do que as mais complexas. Nas árvores de decisão complexas, pode ocorrer o chamado "*overfitting*", ou seja, um superajuste do modelo às amostras de treinamento, o que o impede de classificar corretamente as classes que não se enquadram em todos os requisitos previstos na árvore de decisão. Segundo Körting (2012), uma árvore muito grande pode superajustar os dados, enquanto uma muito pequena pode deixar de capturar estruturas importantes. Por isso, o autor diz ser preferível uma árvore média, desde que não subestime ou superestime os dados, e que também seja facilmente interpretada pelo usuário.

As Tabelas 4A e 4B mostram as matrizes de confusão resultantes da classificação da área-teste B com os algoritmos C4.5 no InterIMAGE e *SimpleCart* no WEKA, respectivamente. Os índices Kappa obtidos, de 0,53 e 0,59, podem ser considerados como bom de acordo com a literatura. Nesta área-teste, a classificação com o algoritmo *SimpleCart* (Figura 3) obteve melhor desempenho que o C4.5, apesar de essas diferenças não serem significantes de acordo com o teste z.

Observando-se o Kappa condicional da área-teste B obtido nas duas metodologias, percebe-se que nesta área ocorreram maiores confusões entre os estádios sucessionais da vegetação, inclusive no estádio inicial. No entanto, a vegetação desta área é mais complexa, apresentando maior heterogeneidade nos diferentes estádios, além de estarem intrinsecamente conectadas e sem influências antrópicas diretas, diferentemente da outra área. Percebe-se que a VEI tem maior variabilidade espectral nesta área-teste, pois possui alguns locais onde predomina uma vegetação arbustiva, e em outros, de gramíneas. A VEM teve várias amostras classificadas como VEI em ambos os métodos. Além de esta classe ser menos representativa na imagem, a VEM é um estádio de transição entre o inicial e avançado, apresentando comportamento espectral semelhante a essas duas classes.

Tabela 4: Matriz de confusão da classificação da área-teste B usando o InterIMAGE com o algoritmo C4.5 (A); e no WEKA e QGIS com o algoritmo *SimpleCart* (B).

		Mapa de referência						
	Classes	VEI	VEM	VEA	Campo	Sombra	TCl	Kcu
Mapa classificado (A)	VEI	130	18	46	12	3	209	0,41
	VEM	16	29	24	0	0	69	0,36
	VEA	33	1	123	1	1	159	0,62
	Campo	1	2	0	37	0	40	0,92
	Sombra	0	0	7	0	16	23	0,68
	TCo	180	50	200	50	20	500	0,53
	Ep	72%	58%	62%	74%	80%	67%	
	Eu	62%	42%	77%	92%	70%		
Mapa classificado (B)	VEI	133	17	35	4	0	189	0,54
	VEM	17	29	18	0	0	64	0,39
	VEA	24	1	126	0	2	153	0,71
	Campo	4	2	0	46	0	52	0,87
	Sombra	2	1	21	0	18	42	0,40
	TCo	180	50	200	50	20	500	0,59
	Ep	74%	58%	63%	92%	90%	70%	
	Eu	70%	45%	82%	88%	43%		

TCo= total de amostras coletadas; TCl: total de amostras classificadas; Ep= Exatidão do produtor; Ep= Exatidão do usuário; Kcu= Kappa condicional do usuário.

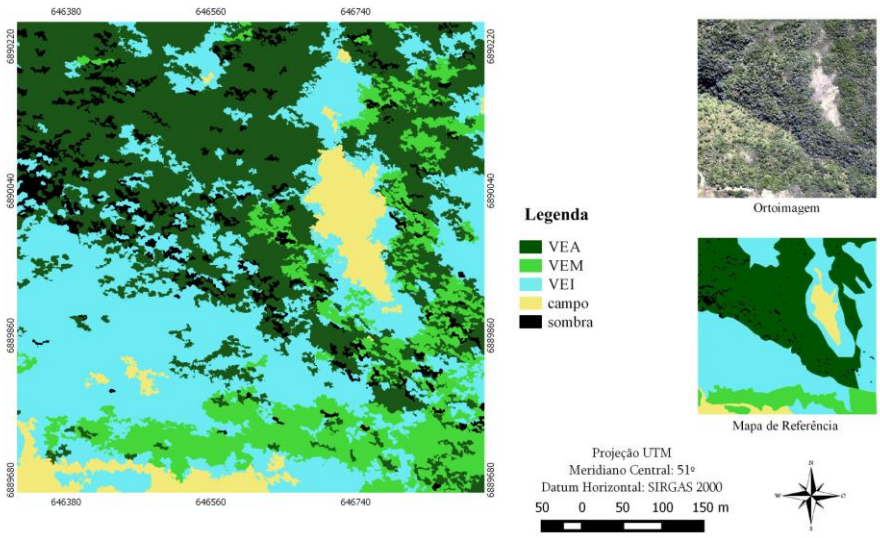

Figura 3: Mapa de uso e cobertura da terra gerado com o algoritmo *SimpleCart* do WEKA.

A Figura 4 ilustra os principais erros de classificação ocorridos nesta área-teste. Na parte superior da figura, percebem-se áreas de VEA classificadas como VEI. Trata-se de locais em que a vegetação, mesmo sendo de grande porte, possui maior brilho e, consequentemente, maiores valores da média dos ND das bandas do visível, característica dos estádios iniciais de regeneração. Na parte inferior, são áreas de estádio inicial que, por possuírem padrão espectral mais escuro e maior heterogeneidade, foram classificadas como VEA.

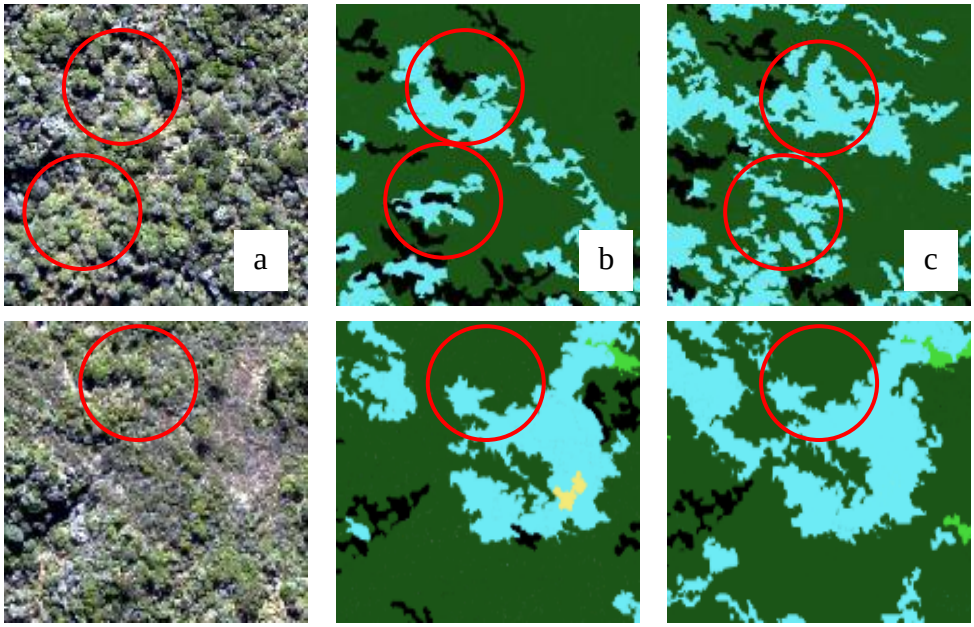

Figura 4: Principais confusões ocorridas na classificação da área-teste C; (a) imagem original; (b) classificação com o *SimpleCart*; (c) classificação com o C4.5. Os círculos mostram áreas em que ocorreu confusão.

A classe sombra manteve baixos índices de exatidão do usuário e Kappa condicional do usuário nas duas áreas testadas. Porém, cabe destacar que o fotointérprete não considerou as sombras menores no interior da vegetação na elaboração do mapa de referência, apenas aquelas mais representativas. Por este motivo, na maioria das metodologias, essa classe teve uma superclassificação em relação ao mapa de referência, e, portanto, maiores erros de comissão.

Os resultados deste estudo ainda podem ser aprimorados aplicando-se uma heurística para reduzir erros de comissão resultantes da classificação automática, editando-se os limiares da árvore, por exemplo, e adequando-se os resultados conforme o mapa de referência. Isto constitui uma vantagem ao se utilizar algoritmos de árvore de decisão, pois, além de permitir que os limiares sejam alterados, de forma a reduzir os erros encontrados na classificação, também permite que os modelos obtidos sejam ajustados às peculiaridades da cena em análise e replicados em outras áreas com características semelhantes de uso e cobertura da terra (Camargo et al., 2009).

4. Conclusão

As técnicas de mineração de dados permitiram constatar que a vegetação apresenta características espectrais e texturais distintas conforme o seu estádio sucessional da vegetação. A etapa da seleção de atributos atendeu ao propósito de redução da dimensão e aumento da acurácia dos modelos de árvore de decisão gerados. Em geral, a componente MDS contribuiu significativamente para a classificação de uso e cobertura da terra nas duas áreas-teste, da mesma forma que as operações entre as bandas correspondentes à região do infravermelho próximo e vermelho.

Os resultados das classificações com os algoritmos C4.5 e *SimpleCart* não apresentaram diferenças significantes para as duas áreas testadas. Porém, o algoritmo *SimpleCart* aliado à etapa de seleção de atributos conseguiu bons resultados com árvores menores e utilizando menos atributos que o C4.5.

A árvore de decisão possui a vantagem de permitir ao usuário visualizar o processo de classificação. Ela pode ser aperfeiçoada de forma a melhorar a acurácia de acordo com as peculiaridades de cada local. A estratégia de uso de uma estrutura hierárquica para a descrição automática das classes com base em algoritmos supervisionados de mineração de dados mostrou-se ser uma forma eficaz e de reduzida subjetividade para a classificação do uso e cobertura da terra. O fato de estas metodologias terem envolvido procedimentos automáticos faz com que sejam aplicáveis em outras áreas. Desta forma, a classificação automatizada é uma alternativa interessante para o estudo de áreas extensas, economizando assim esforço do operador bem como tempo de processamento computacional.

Este trabalho demonstrou que as metodologias testadas são aplicáveis em outras áreas do Bioma Mata Atlântica, além de ter contribuído para uma avaliação comparativa das diversas ferramentas disponíveis gratuitamente para a interpretação de ortoimagens disponíveis no sítio sigsc.sds.sc.gov.br.

REFERÊNCIAS BIBLIOGRÁFICAS

Achanccaray Diaz, P. M. "A Comparison of Segmentation Algorithms for Remote Sensing." *Diss.*, Pontifícia Universidade Católica do Rio de Janeiro, 2014.

Amaral, M. V. F.; Souza, A. L. de; Soares, V. P.; Soares, C. P. B.; Leite, H. G.; Martins, S. V.; Filho, E. I. F.; Lana, J. M. "Avaliação e comparação de métodos de classificação de imagens de satélites para o mapeamento de estádios de sucessão florestal," *Revista Árvore* 33 (2009): 575-82.

Baatz M.; Schäpe, A. "Multiresolution segmentation — An optimization approach for high quality multi-scale image segmentation," *Angewandte Geographische Informations verarbeitung XII.* Karlsruhe, Germany, p. 12−23, 2000.

Blaschke, T.; Strobl, J. "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS," *GIS–Zeitschrift für Geoinformationssysteme* 14 (2001): 12–17, 2001.

Brasil. "Lei n. 11.428, de 22 de dezembro de 2006. Dispõe sobre a utilização e proteção da vegetação nativa do Bioma Mata Atlântica, e dá outras providências." Publicada no *Diário Oficial da União* Seção 1 em 26 de dezembro de 2006. p. 1.

Brasil. Conselho Nacional do Meio Ambiente. 1994. "Resolução CONAMA nº 04/94, de 4 de maio de 1994." Publicada no *Diário Oficial da União* em 17 jun. 1994, n. 114.

Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees.* Belmont, CA: Wadsworth, 1984.

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

Congalton, R. G. and Green, K. *Assessing the accuracy of remotely sensed data: principles and practices*. New York: Lewis Publishers, 1999.

Costa, G. A. O. P., Pinho, C. M. D.; Feitosa, R. Q.; Almeida, C. M.; 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," *Revista Brasileira de Cartografia* 60 (2008): 331-337.

Engemap Geoinformação. "Relatório de produção final" - edital de concorrência pública n. 0010/2009. Florianópolis SC, 218 p., 2012.

Faxina, T. C. "Dilemas da regularização fundiária amigável no Parque Nacional de São Joaquim: Um estudo de caso – a valorização de áreas silvestres." *Diss.*, Universidade do Estado de Santa Catarina, 2014.

Feitosa, R. Q. ; Costa, G. A. O. P., Cazes, T. B.; Feijo, B. "A Genetic Approach for the Automatic Adaptation of Segmentation Parameters" paper presented at 1st International Conference on Object Based Image Analysis, July, Salzburg, 2006.

Feitosa, R. Q.; Costa, G. A. O. P.; Fredrich, C. M. B.; Camargo, F. F.; Almeida, C. M. "Uma Avaliação de Métodos Genéticos para Ajuste de Parâmetros de Segmentação". In: XIV Simpósio Brasileiro de Sensoriamento Remoto - XIV SBSR, 2009, Natal, RN. *Anais do XIV Simpósio Brasileiro de Sensoriamento Remoto*. São José dos Campos, SP: Instituto Nacional de Pesquisas Espaciais, INPE, 2009. p. 6875-6882.

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.* 18 (2012): 302-326.

Gao, Y.; Mas, J.F.; Kerle, N.; Navarrete Pacheco, J. A. "Optimal region growing segmentation and its effect on classification accuracy." *Int. J. Remote Sens.* 32 (2011): 3747–3763.

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

Instituto Brasileiro de Geografia e Estatística. "Informações das cidades do Brasil.*"* Acessado em 12 de abril, 2015. http://www.cidades.ibge.gov.br.

Karegowda, A. G.; Manjunath, A. S.; Jayaram, M. A. "Comparative of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management,* 2 (2010): 271-277.

Klein, R. M. *Mapa fitogeográfico do Estado de Santa Catarina.* Itajaí: Herbário Barbosa Rodrigues; Florianópolis: Universidade Federal de Santa Catarina, 1978.

Körting, T. S. "GeoDMA: a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery." *PhD thesis,* Instituto Nacional de Pesquisas Espaciais, 2012.

Leonardi, F.; Almeida, C. M.; Fonseca, L. M. G. "An ALTM Digital Height Model Associated with VHR Imagery for the Object-Based Classification of Intra-Urban Targets". In: Joint Urban Remote Sensing Event, 2013, São Paulo, SP. *Proceedings of the Joint Urban Remote Sensing Event - JURSE 2013*. Piscataway, NJ, EUA: IEEE Geoscience and Remote Sensing Society, 2013.

Novack, T. "Classificação da cobertura da terra e do uso do solo urbano utilizando o sistema InterIMAGE e imagens do sensor." *Diss.*, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2009.

Piazza, G. A. "Processamento digital de imagens de alta resolução espacial com enfoque na classificação dos estágios sucessionais iniciais da Floresta Ombrófila Densa em Santa Catarina." *Diss.*, Universidade Regional de Blumenau, 2014.

Ponzoni, F. J.; Shimabukuro, Y. E.; Kuplich, T. M. *Sensoriamento Remoto da Vegetação*. São Paulo: Oficina de Textos, 2012.

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

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* 142 (2009): 1141-1153.

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." *Diss*., Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Vibrans, A. C.; Sevegnani, L.; Gasper, A. L.; Lingner, D. V. *Inventário Florístico Florestal de Santa Catarina:* Diversidade e conservação dos remanescentes florestais. Blumenau: Edifurb, 2012.

Vieira, I. C. G.; Almeida, A. S.; Davidson, E. A.; Stone, T. A.; Carvalho, C. J. R.; Guerrero, J. B. "Classifying successional forests using Landsat spectral properties and ecological characteristics in eastern Amazônia." *Remote Sensing of Environment* 87 (2003): 470-481.

Yu, Q.; Gong, P.; Clinton, N.; Biging, G.; Kelly, M.; Schirokauer, D. "Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery." *Photogrammetric Engineering & Remote Sensing* 72 (2006): 799–811.