Ministério da
Ciência, Tecnologia e Inovação

sid.inpe.br/mtc-m21b/2014/02.25.03.03-TDI

MAPEAMENTO DE FORMAÇÕES CAMPESTRES NATIVAS E DE PASTAGENS CULTIVADAS NO CERRADO BRASILEIRO UTILIZANDO MINERAÇÃO DE DADOS

Wanderson Santos Costa

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Leila Maria Garcia Fonseca, e Thales Sehn Körting, aprovada em 21 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FQP36L>

INPE
São José dos Campos
2014

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

Ministério da
Ciência, Tecnologia e Inovação

sid.inpe.br/mtc-m21b/2014/02.25.03.03-TDI

MAPEAMENTO DE FORMAÇÕES CAMPESTRES NATIVAS E DE PASTAGENS CULTIVADAS NO CERRADO BRASILEIRO UTILIZANDO MINERAÇÃO DE DADOS

Wanderson Santos Costa

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Leila Maria Garcia Fonseca, e Thales Sehn Körting, aprovada em 21 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FQP36L>

INPE
São José dos Campos
2014

Dados Internacionais de Catalogação na Publicação (CIP)

Costa, Wanderson Santos.

C822m Mapeamento de formações campestres nativas e de pastagens cultivadas no Cerrado brasileiro utilizando mineração de dados / Wanderson Santos Costa. – São José dos Campos : INPE, 2014.

xix + 85 p. ; (sid.inpe.br/mtc-m21b/2014/02.25.03.03-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Orientadores : Drs. Leila Maria Garcia Fonseca, e Thales Sehn Körting.

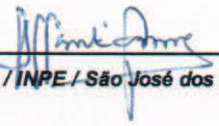
1. Cerrado brasileiro. 2. processamento de imagens. 3. mineração de dados. I.Título.

CDU 528.854(213.54)

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de ***Mestre*** em

Computação Aplicada

Dr. Sidnei João Siqueira Sant'Anna



Presidente / INPE / São José dos Campos - SP

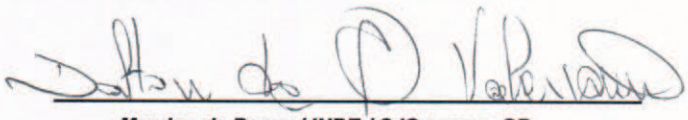
Dra. Leila Maria Garcia Fonseca

Orientador(a) / INPE / SJCampos - SP

Dr. Thales Sehn Korting

Orientador(a) / INPE / São José dos Campos - SP

Dr. Dalton de Morisson Valeriano



Membro da Banca / INPE / SJCampos - SP

Dr. João Camargo Neto

Convidado(a) / EMBRAPA / Campinas - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): ***Wanderson Santos Costa***

São José dos Campos, 21 de Fevereiro de 2014

A meus pais, Valdenira e Valdercio.

AGRADECIMENTOS

Aos meus orientadores Leila Maria Garcia Fonseca e Thales Sehn Körting, pela competência e dedicação. Agradeço pela atenção e confiança depositadas.

A meus pais e meu irmão, pelo incentivo e apoio singulares.

Aos amigos de Itabaiana e Aracaju, pelos longos anos de amizade e por entenderem meus momentos de ausência, em especial à Tâmara. Obrigado pelo apoio constante e pela força nos momentos difíceis.

Aos amigos que fiz nestes dois anos de pós-graduação, pelo companheirismo e conhecimentos compartilhados.

Ao Instituto Nacional de Pesquisas Espaciais – INPE, pela oportunidade de estudo e utilização de suas instalações.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, pelo auxílio financeiro.

RESUMO

O Cerrado é o segundo maior bioma do território brasileiro, com demandas conflitantes de preservação ambiental e atividades agropastoris. Dentre as mudanças do uso e cobertura do solo no Cerrado, mais de 500.000 km² do bioma foram transformados em áreas de pastagem cultivada nos últimos anos. A discriminação entre as formações vegetais nativas e a identificação dos tipos de uso e cobertura no Cerrado são informações importantes para a política de proteção e monitoramento deste bioma. Neste trabalho, foi desenvolvida uma metodologia baseada em técnicas de sensoriamento remoto, incluindo integração de dados multitemporais e de múltiplas resoluções, e mineração de dados, para mapear áreas de pastagem cultivada e áreas de vegetação nativa correspondente às formações campestres no Cerrado brasileiro. Dados referentes ao uso e cobertura do solo, relevo, informações espectrais de imagens Landsat, índices de vegetação e componentes de solo e sombra extraídos do modelo de mistura espectral foram utilizados no processo de classificação. Os resultados mostraram que, por meio dos algoritmos de árvores de decisão, SVM (*Support Vector Machines*) e florestas aleatórias, a análise e integração de dados auxiliou na classificação da região de interesse. Ao discriminar áreas de pastagem cultivada e campo nativo, obteve-se uma taxa de acerto de cerca de até 87% na área de estudo, localizada em Minas Gerais, sendo então possível identificar atributos, regras e dados necessários para reconhecer, por meio de imagens de sensoriamento remoto, áreas campestres nativas e plantadas no Cerrado brasileiro.

MAPPPING NATIVE GRASSLANDS FORMATIONS AND CULTIVATED PASTURES IN THE BRAZILIAN CERRADO USING DATA MINING

ABSTRACT

Cerrado is the second largest biome in Brazil. Among the use and land cover changes in the Cerrado, over 500,000 km² of the biome have been transformed into cultivated pastures in recent years. Distinguishing the native formations types and the identification of land use and land cover types in the Cerrado are important for monitoring and defining protection policy of the biome. Within this context, this work aims at developing a methodology based on remote sensing techniques, including multiresolution and multitemporal data fusion, and data mining, to map pasture and native grassland areas in the Brazilian Cerrado. Data related to land use and cover, relief, spectral information from Landsat images, vegetation indices, soil and shade components extracted from the Linear Spectral Mixture Model were used to perform the image classification. Decision trees, Support Vector Machines and Random Forests algorithms were used, and the results showed that the analysis and integration of different data sources can aid in the classification process. In order to discriminate areas of cultivated pastures and grassland formations, we obtained accuracies up to 87% in the study area, located in Minas Gerais state, being able to identify attributes, rules and data required to recognize these areas in the Brazilian Cerrado by remote sensing images.

LISTA DE FIGURAS

Pág.

Figura 1.1 Distribuição espacial das classes de uso da terra no bioma Cerrado referente ao ano de 2002. 1
Figura 2.1. Mapa de biomas brasileiros. 5
Figura 2.2. Fitofisionomias do bioma Cerrado. 7
Figura 2.3. Regiões de Campo Limpo. 11
Figura 2.4. Aspecto de áreas de Campo Sujo. 11
Figura 2.5. Região de Campo Rupestre. 12
Figura 2.6. Exemplos de pasto manejado (à esquerda) e degradado (à direita). 13
Figura 3.1. Exemplo de série temporal do índice de vegetação no pixel (*i,j*). .. 16
Figura 3.2. Visão simplificada do domínio do problema de análise temporal. .. 18
Figura 3.3. Intervalos de eventos no tempo. 19
Figura 3.4. Exemplo da aplicação do MLME e suas imagens-fração resultantes. 22
Figura 4.1. Etapas do KDD. A mineração de dados corresponde a uma das etapas do processo da transformação de dados em conhecimento. 25
Figura 4.2. Exemplo de árvore de decisão indicando se uma região é uma provável floresta ou um campo. 29
Figura 4.3. Exemplo do processo de seleção de uma classe a partir da classificação em floresta aleatória. 35
Figura 4.4. Exemplos de retas capazes de separar as classes -1 e +1. Há um número infinito de retas que separam as duas classes. 37
Figura 4.5. Representação do hiperplano ótimo e dos vetores de suporte. 38
Figura 4.6. Exemplo de mapeamento para um espaço de maior dimensão, fornecendo uma separação linear. 41
Figura 5.1 Área de estudo do trabalho, localizada no sudoeste do estado de Minas Gerais. Recorte de uma imagem *Landsat* TM-5 (R5B4G3) da região. .. 45
Figura 5.2 Etapas de processamento e análise dos dados. 46
Figura 5.3 Mapa de referência utilizado na área de estudo para obtenção das regiões de formações campestres: Campo (cor ciano) e Campo Rupestre (cor rosa). 47
Figura 5.4 Regiões do Cerrado mapeadas pelo MMA como áreas naturais (em verde) e antrópicas (em amarelo). 48
Figura 5.5 Imagens TOPODATA relativas à altitude (esquerda) e à declividade (direita) da área de estudo. 50
Figura 5.6 Imagens *Landsat* TM-5 (R5B4G3) de 2006 (esquerda) e 2009 (direita). 51
Figura 5.7 Sequência de imagens para compor os perfis anuais de EVI e EVI2 para o ano de 2009. 52

Figura 5.8 Componentes do MLME relativos ao ano de 2006..... 53
Figura 5.9 Abordagem de pixels puros..... 54
Figura 5.10 Exemplo de polígono desconsiderado do conjunto de amostras na abordagem de pixels puros. 55
Figura 6.1 Árvore de decisão para o Experimento 1 (EVI + Landsat + TOPODATA). 63
Figura 6.2. Árvore de decisão (à esquerda) e mapeamento (à direita) para o melhor resultado encontrado..... 69

LISTA DE TABELAS

Pág.

Tabela 2.1 – Principais usos do solo no Cerrado. 6
Tabela 4.1 – Amostras de treinamento para uma base de dados de cobertura do solo..... 31
Tabela 6.1 – Resumo dos conjuntos de dados utilizados em cada experimento. 58
Tabela 6.2 – Resultados do Experimento 1 para 4 classes. 59
Tabela 6.3 – Matriz de confusão para o Experimento 1 (EVI com florestas aleatórias)..... 61
Tabela 6.4 – Resultados do Experimento 1 para 3 classes. 62
Tabela 6.5 – Resultados do Experimento 2 para 4 classes. 64
Tabela 6.6 – Resultados do Experimento 2 para 3 classes. 65
Tabela 6.7 – Resultados do Experimento 3 para 4 classes. 67
Tabela 6.8 – Resultados do Experimento 3 para 3 classes. 67
Tabela 6.9 – Resultados do Experimento 4 para 3 e 4 classes..... 68
Tabela 6.10 – Matriz de confusão do Experimento 4 (3 classes – Florestas aleatórias)..... 70
Tabela 6.11 – Comparação dos resultados mais relevantes (3 Classes)..... 70
Tabela 6.12 – Comparação dos resultados mais relevantes (4 Classes)..... 72

LISTA DE SIGLAS E ABREVIATURAS

EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EVI	*Enhanced Vegetation Index*
EVI2	*Enhanced Vegetation Index 2*
FMC	*Fuzzy Markov Chain*
GeoDMA	*Geographic Data Mining Analyst*
HMM	*Hidden Markov Model*
IBGE	Instituto Brasileiro de Geografia e Estatística
MLME	Modelo Linear de Mistura Espectral
MODIS	*Moderate Resolution Imaging Spectroradiometer*
NDVI	*Normalized Difference Vegetation Index*
OPF	*Optimum Path Forest*
PCA	*Principal Component Analysis*
RNA	Rede Neural Artificial
RWFR	*Renewable Water Fresh Resources*
SAVI	*Soil-Adjusted Vegetation Index*
SVM	*Support Vector Machine*

SUMÁRIO

Pág.

1. INTRODUÇÃO 1
1.1. Organização do trabalho 4
2. O CERRADO BRASILEIRO 5
2.1. Fitofisionomias do Cerrado 7
2.1.1. Formações campestres 10
2.2. Pastagens cultivadas 12
3. PROCESSAMENTO E ANÁLISE DE IMAGENS 15
3.1. Técnicas de detecção de mudanças 18
3.1.1. Álgebra 20
3.1.2. Transformação 20
3.1.2.1. Modelo linear de mistura espectral 20
3.1.3. Interpretação visual 23
3.1.4. Classificação de imagens 23
4. MINERAÇÃO DE DADOS 25
4.1. Classificação usando árvores de decisão 28
4.2. Classificação usando florestas aleatórias 34
4.3. Classificação usando SVM 36
5. METODOLOGIA 45
5.1. Mapas de referência 47
5.2. Dados auxiliares 49
5.3. Imagens *Landsat* 50
5.4. Imagens MODIS 51
5.5. Transformação 52
5.6. Extração de atributos 53
5.7. Treinamento e classificação 56
5.8. Avaliação 56
6. RESULTADOS E DISCUSSÕES 58
7. CONCLUSÕES E TRABALHOS FUTUROS 73
REFERÊNCIAS BIBLIOGRÁFICAS 75

1. INTRODUÇÃO

Fatores como crescimento populacional, mudanças climáticas e a contínua demanda de energia, de água e de alimentos constituem uma ameaça de caráter global, podendo ocasionar sérios riscos ambientais caso os recursos naturais não sejam adequadamente utilizados (BEDDINGTON, 2009). Dentro desta perspectiva, pode-se mencionar a questão das mudanças no uso e cobertura do solo do segundo maior bioma brasileiro: o Cerrado (RATTER *et al*., 1997). Mais da metade da área do Cerrado brasileiro tem sido transformada, principalmente, em áreas de pastagem e agricultura, perdendo cerca de 1 milhão de km² de sua vegetação original (MACHADO *et al*., 2004). As áreas de agricultura cobrem mais de 100.000 km² e as de pastagem superam 500.000 km², enquanto que as áreas preservadas envolvem apenas cerca de 33.000 km² (KLINK; MACHADO, 2005). Uma distribuição das áreas transformadas no Cerrado pode ser observada na Figura 1.1.

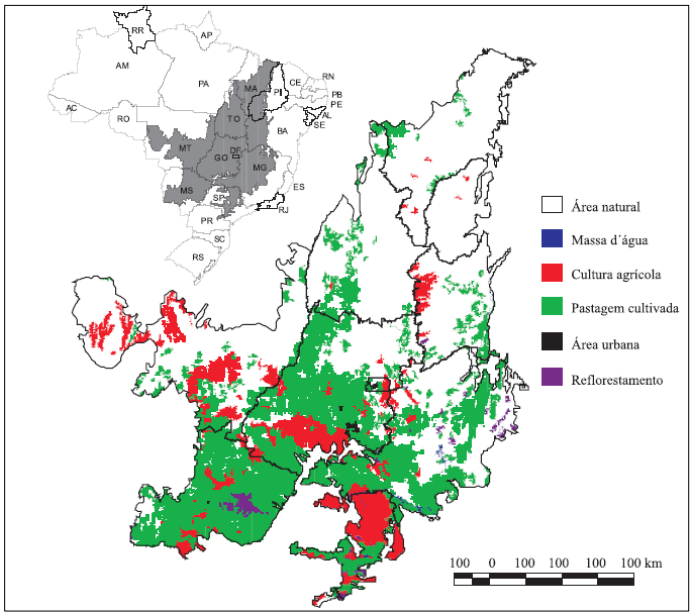


Figura 1.1 Distribuição espacial das classes de uso da terra no bioma Cerrado referente ao ano de 2002.

Fonte: Sano *et al*. (2008)

A destruição das formações florestais, savânicas e campestres do Cerrado é três vezes maior que a quantidade da área desflorestada da região Amazônica. A área desmatada por ano apresenta uma taxa ainda mais alta do que se observa na Amazônia (KLINK; MACHADO, 2005; SANO *et al.*, 2008). Esta situação representa um alto custo ambiental, ocasionando perda de biodiversidade, erosão do solo, degradação das formações vegetais do bioma, poluição da água, mudanças nos eventos de fogo típicos do bioma, instabilidade do ciclo do carbono e prováveis modificações climáticas regionais (KLINK; MACHADO, 2005).

O Cerrado tem recebido menos atenção que a Amazônia em termos de medidas de preservação ambiental (SCARIOT *et al.*, 2005), de forma que apenas cerca de 2% da área é legalmente protegida. Entretanto, as transformações no uso e cobertura do solo do Cerrado podem também influenciar de forma significativa as mudanças nas condições climáticas regionais na floresta Amazônica (MALHADO *et al*. 2010). Costa e Pires (2010) mostram que os impactos climáticos do desmatamento da floresta Amazônica, combinados com o desflorestamento contínuo do Cerrado, podem produzir um aumento substancial na duração da estação seca em regiões ao sudeste da Amazônia.

Segundo Walter (2006), o Cerrado pode ser expresso, de acordo com o senso comum, como uma “paisagem com um estrato gramíneo contínuo (ou descontínuo), contendo árvores ou arbustos”. Existe, contudo, um grande número de definições sobre Cerrado e, em decorrência, várias propostas de classificação das formações vegetais no bioma. Dentre as formações vegetais existentes no bioma, encontram-se as formações campestres, que se referem a regiões com predominância de espécies herbáceas e algumas arbustivas, sem ou com pouca ocorrência de árvores na paisagem (RIBEIRO; WALTER, 2008; IBGE, 2012).

As vegetações incluídas nas formações campestres são áreas de Campo Sujo, Rupestre e Limpo. Regiões com Campo Sujo e Campo Rupestre apresentam um tipo fisionômico predominantemente herbáceo-arbustivo. Contudo, as áreas de Campo Rupestre agrupam paisagens em micro relevos com espécies típicas, geralmente, ocupando trechos de afloramento rochosos em altitudes superiores a 900 m. Por outro lado, na fitofisionomia de Campo Limpo têm-se a predominância de gramíneas entremeadas com algumas plantas lenhosas raquíticas com ausência de árvores (RIBEIRO; WALTER, 2008).

Considerando que estas fitofisionomias são formações naturais, elas são passíveis de recuperação caso sejam degradadas. Neste caso, o monitoramento deste tipo de vegetação, por meio de imagens de satélite, pode auxiliar na recuperação da sua integridade física, química e biológica assim como na capacidade produtiva da região degradada (RODRIGUES; GANDOLFI, 2001). Assim, mapas de uso e cobertura do Cerrado, obtidos a partir imagens de satélites e técnicas de sensoriamento remoto, podem ser usados para monitorar estas áreas.

Existem alguns estudos, como o de Ferreira *et al*. (2013b), que usam técnicas de análise multitemporal baseada em índices de vegetação para classificar áreas do Cerrado. Entretanto, as áreas de pastagem antropizada, de forma similar às formações campestres supracitadas, podem variar de formações com predominância de gramíneas a locais que apresentam dominância de espécies arbustivas, por exemplo (EMBRAPA; INPE, 2011). Desta forma, o mapeamento de áreas de Pastagem e de formações campestres nativas no Cerrado, por meio de imagens de satélites, pode se tornar uma tarefa difícil quando usada apenas a informação espectral dos alvos (SANO *et al.*, 2008).

Dentro deste contexto, este trabalho tem como objetivo desenvolver uma metodologia para mapeamento de áreas de Pastagem Cultivada e Campo Nativo (Limpo, Sujo e Rupestre) utilizando técnicas de integração de dados em

múltiplas resoluções espaciais e espectrais, análise multitemporal de imagens de satélite e mineração de dados.

1.1. Organização do trabalho

O trabalho foi organizado da seguinte forma: no Capítulo 2 é apresentada uma breve descrição do bioma Cerrado brasileiro, com ênfase nas formações vegetais e áreas antrópicas de interesse. Nos Capítulos 3 e 4 são apresentadas as técnicas de processamento e análise de imagens usadas neste estudo. O procedimento metodológico usado para produzir e avaliar as imagens classificadas é apresentado no Capítulo 5. No Capítulo 6 é apresentada uma discussão sobre os progressos obtidos neste estudo. Finalmente, no Capítulo 7 são apresentadas as considerações finais e trabalhos futuros.

2. O CERRADO BRASILEIRO

Uma das regiões de maior biodiversidade do planeta, o Cerrado brasileiro apresenta uma área de aproximadamente 2 milhões de km², compreendendo cerca de 24% do território nacional, superado em área apenas pela Amazônia (BROSSARD; BARCELLOS, 2005). O termo Cerrado é comumente empregado para indicar o conjunto de ecossistemas (savanas, campos e matas) que ocupa a região central do Brasil (EITEN, 1977), estendendo-se desde o litoral nordeste do estado do Maranhão até o norte do estado do Paraná, como pode ser observado na Figura 2.1.

Figura 2.1. Mapa de biomas brasileiros.
Fonte: IBGE (2004)

Segundo Ferreira *et al*. (2003), o Cerrado brasileiro corresponde a uma das savanas mais úmidas do planeta, apresentando um clima sazonal, com estações seca e chuvosa bem definidas. A estação seca ocorre no outono/inverno, entre os meses de abril a setembro, enquanto a chuvosa é predominante de outubro a março (KLINK; MACHADO, 2005). Com uma rica biodiversidade, o Cerrado contém aproximadamente 160.000 espécies de plantas, animais e fungos (FERREIRA *et al.*, 2003). Além disso, sua vegetação

é caracterizada pelo predomínio de gramíneas, geralmente intercaladas por árvores e arbustos de tamanhos variados (RIBEIRO; WALTER, 2008).

De acordo com Machado *et al*. (2004), mais da metade da vegetação original do Cerrado foi transformada em áreas de pastagem, agricultura e outros usos, como mostra a Tabela 2.1. Além disso, estudos mostram que as mudanças do uso do solo no Cerrado ocorrem com maior intensidade do que na região Amazônica (SANO *et al*., 2001; SKOLE *et al*., 2012).

Tabela 2.1 – Principais usos do solo no Cerrado.

Uso do solo	**Área (*ha*)**	**Percentual da área**
Áreas nativas	70.581.182	44,53
Pastagem plantada	65.874.145	41,56
Agricultura	17.984.719	11,35
Área de reflorestamento	116.760	0,07
Áreas urbanas	3.006.830	1,90
Outros	930.304	0,59
Total	158.493.921	

Fonte: Machado *et al*. (2004)

Em decorrência das mudanças de uso do solo, surgem ameaças substanciais aos ecossistemas e às espécies do bioma. Segundo Klink e Machado (2005), somente 2,2% de sua área estão sobre proteção legal e diversas espécies de animais e plantas estão ameaçadas de extinção. Além disso, um número estimado de 20% das espécies endêmicas ameaçadas já não existem em áreas preservadas. Além da degradação da vegetação e erosão do solo, a introdução e invasão de espécies exóticas e o uso do fogo para criação de áreas de pastagens são, também, ameaças à biodiversidade do Cerrado.

2.1. Fitofisionomias do Cerrado

De forma geral, o Cerrado pode ser entendido como uma formação composta por campos graminosos coexistentes com árvores e arbustos esparsos. Contudo, várias definições do termo Cerrado e propostas de classificação das formações vegetais do bioma são apresentadas na literatura (WALTER, 2006). Ribeiro e Walter (2008), por exemplo, descrevem o Cerrado como uma região composta de formações florestais, savânicas e campestres.

As formações florestais representam áreas com predominância de espécies arbóreas com formação de dossel. As formações savânicas designam regiões com árvores e arbustos esparsos sobre um campo de gramínea, sem a formação de dossel contínuo. Por outro lado, as formações campestres referem-se a regiões com predominância de espécies herbáceas e algumas arbustivas, sem a ocorrência de árvores na paisagem.

Estas três formações paisagísticas, representadas na Figura 2.2, são classificadas em diferentes fitofisionomias, baseadas primeiramente na forma (fisionomia), que é definida pela estrutura, pelas formas de crescimento dominantes e por possíveis mudanças estacionais. Posteriormente, consideram-se aspectos do ambiente e da composição florística.

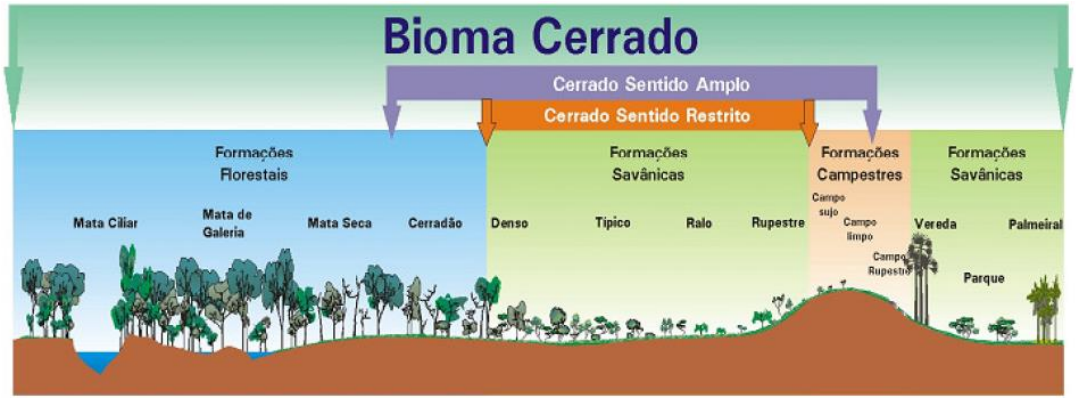


Figura 2.2. Fitofisionomias do bioma Cerrado.

Fonte: Ribeiro e Walter (2008).

Ribeiro e Walter (2008) descrevem os tipos fitofisionômicos da seguinte forma:

1. Formações Florestais: englobam os tipos de vegetação com predominância de espécies arbóreas, com a formação de dossel contínuo.

    a. Mata Ciliar: vegetação florestal que acompanha os rios de médio porte da Região do Cerrado, em que a vegetação arbórea não forma galerias. É, em geral, relativamente estreita e apresenta considerável proporcionalidade entre a largura em cada margem e o leito do rio, dificilmente ultrapassando 100 m de largura em cada margem.

    b. Mata de Galeria: vegetação florestal perenifólia que acompanha os rios de pequeno porte e os córregos dos planaltos do Brasil Central, formando corredores fechados sobre o curso de água. Localiza-se, geralmente, nos fundos dos vales ou em cabeceiras de drenagem onde os cursos de água ainda não escavaram um canal definitivo.

    c. Mata Seca: caracterizada por diversos níveis de caducifólia durante a estação seca, a vegetação é encontrada em interflúvios, em locais geralmente mais ricos em nutrientes. Engloba as formações florestais no bioma que não possuem associação com cursos de água.

    d. Cerradão: formação florestal do bioma Cerrado com vegetação que apresentam folhas duras e coriáceas, motivo pelo qual é incluído no limite mais alto do conceito de Cerrado sentido amplo. Pode ser caracterizado como um sub-bosque formado por pequenos arbustos e ervas, com poucas gramíneas, com presença dominante de espécies que ocorrem no Cerrado sentido restrito e também por espécies de florestas.

2. Formações Savânicas: englobam quatro tipos fitofisionômicos principais: o Cerrado sentido restrito, o Parque de Cerrado, o Palmeiral e a Vereda.

    a. Cerrado Sentido Restrito: destaca-se pela presença dos estratos arbóreo e arbustivo-herbáceo definidos, com árvores aleatoriamente distribuídas sobre o terreno sem formar um dossel contínuo, em diferentes densidades. De acordo com a densidade (estrutura) arbóreo-arbustiva, ou com o ambiente em que se encontra, o Cerrado sentido restrito apresenta quatro subtipos: Cerrado Denso, Cerrado Típico, Cerrado Ralo e Cerrado Rupestre. O Cerrado sentido restrito caracteriza-se pela presença de árvores baixas, inclinadas, tortuosas, com ramificações irregulares e retorcidas, geralmente com evidências de queimadas. Os arbustos e subarbustos encontram-se espalhados, com algumas espécies apresentando órgãos subterrâneos perenes, que permite a rebrota após a queima ou corte.

    b. Parque de Cerrado: a ocorrência de árvores é concentrada em locais específicos, agrupadas em pequenas elevações do terreno, algumas vezes imperceptíveis e outras com destaque significativo. As árvores possuem altura média de 3 m a 6 m e forma-se uma cobertura arbórea de 5% a 20%.

    c. Palmeiral: pode ocorrer tanto em áreas bem drenadas como em mal drenadas e caracteriza-se pela presença marcante de determinada espécie de palmeira arbórea. Não há destaque de árvores dicotiledôneas, embora elas possam ocorrer com frequência baixa. Podem ser encontrados, no mínimo, quatro subtipos mais comuns de palmeirais no bioma e, pelo domínio de determinada espécie, o trecho de vegetação pode ser atribuído pelo nome comum da palmeira.

d. Vereda: é um tipo de vegetação com a palmeira arbórea buriti emergente, em meio a agrupamentos mais ou menos densos de espécies arbustivo-herbáceas, são circundadas por campos típicos, geralmente úmidos, e não formam dossel. Os buritis mais altos podem chegar a uma altura média de 15 m e a cobertura varia de 5% a 10%.

3. Formações campestres: englobam três tipos de vegetação principais: o Campo Sujo, o Campo Limpo e o Campo Rupestre.

    a. O Campo Sujo caracteriza-se pela presença evidente de arbustos e subarbustos entremeados no estrato arbustivo-herbáceo.

    b. No Campo Limpo, a presença de arbustos e subarbustos é insignificante, com ausência de árvores e predomínio de gramíneas.

    c. O Campo Rupestre possui trechos com estrutura similar ao Campo Sujo, diferenciando-se tanto pelo substrato, composto por afloramentos de rocha, quanto pela composição florística (RIBEIRO; WALTER, 2008).

2.1.1. Formações campestres

Dentre os tipos fitofisionômicos brevemente descritos, este trabalho tem como foco as formações campestres no Cerrado: Campo Limpo, Campo Sujo e Campo Rupestre.

Fitofisionomia com predominância herbácea, o Campo Limpo contém poucos arbustos e há a inexistência de árvores na região (Figura 2.3). É localizado com frequência nas chapadas, olhos d’água e encostas, com locais que podem apresentar diversas posições topográficas e variações no grau de profundidade e fertilidade do solo (RIBEIRO; WALTER, 2008).

Figura 2.3. Regiões de Campo Limpo.
Fonte: IBGE (2012).

Regiões com Campo Sujo contêm, em seu estrato arbustivo-herbáceo, espécies de arbustos e subarbustos muitas vezes compostas por plantas menos desenvolvidas das espécies arbóreas do Cerrado Sentido Restrito. Podem ser encontradas em solos rasos, com presença eventual de pequenos afloramentos de rocha de pouca extensão ou mesmo em terrenos profundos e de pouca fertilidade (RIBEIRO; WALTER, 2008). Áreas de Campo Sujo apresentam variações dependentes de particularidades ambientais, definidas pela topografia e umidade do solo, como ilustradas na Figura 2.4.

Figura 2.4. Aspecto de áreas de Campo Sujo.
Fonte: IBGE (2012).

O tipo fitofisionômico Campo Rupestre, como exposto na Figura 2.5, agrupa paisagens em microrrelevos com espécies típicas, geralmente ocupando trechos de afloramento rochosos em altitudes superiores a 900 metros,

ocasionalmente a partir de 700 metros, com a presença eventual de arvoretas raquíticas. São, em geral, regiões com solos ácidos e pobres em nutrientes, onde existem altas variações de temperatura e ventos constantes (RIBEIRO; WALTER, 2008).

Figura 2.5. Região de Campo Rupestre.
Fonte: Coura (2006).

2.2. Pastagens cultivadas

O efetivo de bovinos no bioma Cerrado representa cerca de 40% do rebanho bovino do país e compreende cerca de 62,7 milhões de cabeças conforme o Censo Agropecuário do IBGE de 2006. Para esta data, existiam aproximadamente 410.000 estabelecimentos agropecuários com pelo menos um bovino, dentre os quais cerca de 100.000 informaram o corte (46,4 milhões de cabeças) como a principal finalidade da criação. Além disso, as áreas de pastagem no Cerrado correspondem a cerca de 500.000 km² (MACHADO *et al*., 2004).

De acordo com Ferreira *et al.* (2013b), é importante realizar a análise da degradação dos pastos cultivados. No mínimo, 50% das áreas de pastagem plantada no Cerrado já se encontram severamente degradadas (Figura 2.6), ocasionando o aumento da erosão e a perda da fertilidade do solo, a redução

da biomassa e a predominância de espécies estrangeiras, principalmente as espécies de gramíneas africanas do gênero *Brachiaria*. Com isso, segundo Chaves *et al*. (2001) a recuperação destas áreas degradadas pode auxiliar no aumento da renda para os produtores e pode-se reduzir o impacto ambiental no Cerrado por meio da diminuição da erosão, da emissão de dióxido de carbono e da abertura de novas áreas para pastagem.

Figura 2.6. Exemplos de pasto manejado (à esquerda) e degradado (à direita).
Fonte: Adaptada de EMBRAPA e INPE (2011).

Dentre os projetos de monitoramento de pastagens cultivadas no Brasil, pode-se destacar o Projeto GeoDegrade (SILVA *et al.* 2013), que está em fase final de desenvolvimento e tem como meta a identificação e monitoramento dos níveis de degradação em pastagens cultivadas em regiões da Amazônia, Cerrado e Mata Atlântica. Este projeto utiliza conjuntos de dados de diferentes escalas e projeções tais como imagens de média e alta resolução, mapas de solo e informações referentes à hidrografia, ao relevo e ao uso e cobertura da terra.

No seu estudo, Ferreira *et al*. (2013b) utilizam dados biofísicos das estações secas e úmidas das áreas de pastagem (tipo de solo, percentual de cobertura vegetal e biomassa), dados radiométricos (NDVI e EVI - *Normalized Difference Vegetation Index* e *Enhanced Vegetation Index*, respectivamente) e climáticos (precipitação e RWFR, *Renewable Water Fresh Resources*). Segundo os autores, as estações seca e úmida são bem definidas na região e, portanto,

podem auxiliar na identificação das características biofísicas e radiométricas das áreas de pastagem.

De acordo com Ferreira *et al*. (2013b), a identificação das regiões de pastagem é difícil porque a degradação dos pastos pode, por exemplo, influenciar na porcentagem da cobertura vegetal e na resposta dos índices de vegetação. Confusão e incertezas podem ocorrer quando as pastagens são manejadas de forma inadequada, uma vez que podem aparecer espécies invasoras ou mesmo o renascimento de espécies de arbustos e árvores nativas nestas regiões.

Em decorrência das grandes extensões de áreas, da periodicidade marcante da vegetação natural e da confusão espectral entre classes, a necessidade de identificar e discriminar áreas de pastagens cultivadas e pastagens naturais (formações campestres) é um grande desafio (SANO *et al.*, 2008). Consequentemente, para melhorar a discriminação destes alvos é necessário usar dados temporais e de campo, e também entender melhor as propriedades biofísicas destas áreas (FERREIRA *et al*., 2013a). Desta forma, a análise usando imagens com diferentes resoluções espaciais (*Landsat*, MODIS, *QuickBird*) e temporais (*AVHRR*), assim como a análise em diferentes intervalos de tempo (por exemplo, mensalmente), podem permitir a identificação dos padrões espaciais e temporais dos alvos no Cerrado de forma sistemática e com mais acurácia.

3. PROCESSAMENTO E ANÁLISE DE IMAGENS

Várias grupos de pesquisa estudam as mudanças do uso e cobertura do solo nos diferentes biomas, e o processamento e a análise de imagens de sensoriamento remoto têm papel fundamental na detecção destas mudanças. A grande quantidade de dados de sensoriamento remoto, combinada com informações dos modelos de ecossistema, oferecem uma boa oportunidade para predição e entendimento do comportamento dos ecossistemas terrestres (TAN *et al*., 2001).

A disponibilidade de uma grande quantidade de dados de sensores remotos ópticos, a exemplo do programa *Landsat*, de sensores de alta resolução espacial como o QuickBird, IKONOS e World-View 2, e outros sensores de média e baixa resolução espacial, tem tornado cada vez mais acessível e de forma mais detalhada e continuada a observação da Terra. O sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) produz dados com alta frequência temporal da superfície terrestre, dos oceanos e da atmosfera. Seus produtos foram projetados para fornecer, de forma consistente, informações espaciais e temporais das condições da vegetação global, que podem ser utilizadas para fazer o monitoramento da cobertura da vegetação (HUETE *et al*., 2002; ZHANG *et al.*, 2003).

A combinação de informações temporais, espaciais e espectrais, quando adequadamente explorada, permite detectar padrões complexos e importantes para o monitoramento e análise da dinâmica da cobertura da terra (BRUZZONE *et al*., 2003). No entanto, a alta resolução dos dados exige uma maior capacidade de processamento para extração e interpretação das informações, além de que outros problemas, tais como falhas nos dados temporais e cobertura por nuvens, podem restringir a utilização dos indicadores de fenologia vegetal para a classificação da cobertura do solo (LAMBIN; LINDERMAN, 2006).

Em estudos de análise e caracterização da cobertura vegetal, utilizam-se índices de vegetação para o monitoramento sazonal e interanual dos parâmetros biofísicos, fenológicos e estruturais da vegetação (HUETE *et al*., 2002). A Figura 3.1 ilustra a criação de uma série temporal de um dado índice de vegetação ($d_{i,j}$). Para cada pixel em questão, pode-se observar uma série temporal que representa a variação da vegetação no decorrer do tempo.

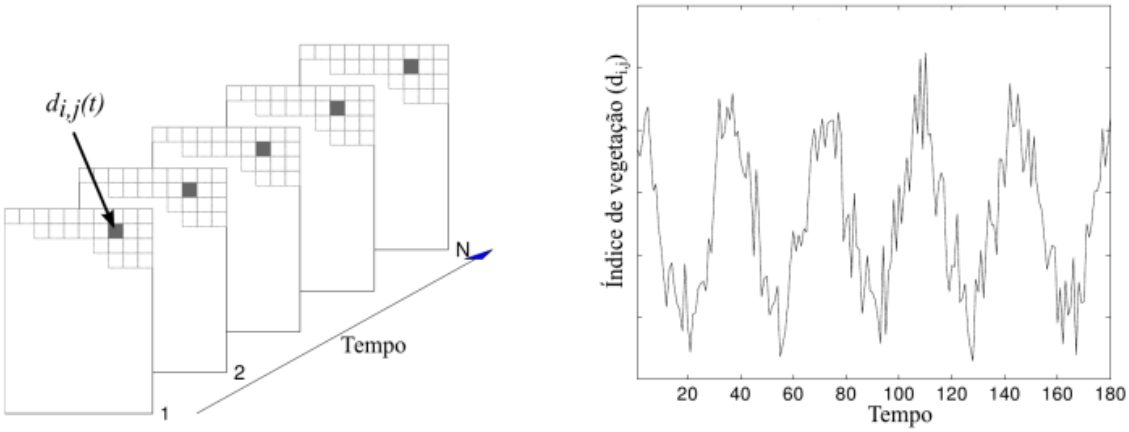


Figura 3.1. Exemplo de série temporal do índice de vegetação no pixel (*i,j*).
Fonte: Adaptada de Eklund e Jönsson (2012).

Segundo Tucker *et al*. (2005), os índices de vegetação produzidos através dos dados MODIS representam medidas aprimoradas das condições espaciais, espectrais e radiométricas da superfície da vegetação. Produzidos geralmente com resoluções espaciais entre 250 m e 1 km, os índices de vegetação do MODIS mais utilizados são o NDVI e EVI (JUSTICE *et al*., 2002). O cálculo do NDVI baseia-se nas reflectâncias dos comprimentos de onda vermelho e infravermelho próximo (TUCKER, 1979):

$$NDVI = \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red}} , \tag{3.1}$$

onde ρ_{NIR} e ρ_{red} são as reflectâncias nas bandas do infravermelho próximo e do vermelho, respectivamente. A razão entre as bandas no cálculo do NDVI

reduz algumas formas de ruído, a exemplo das diferenças de iluminação, das sombras de nuvens e das variações topográficas. Contudo, esse índice possui baixa sensibilidade em regiões com alta concentração de biomassa e pode apresentar limitações relacionadas às variações do brilho do solo (JIANG *et al*., 2008).

Para reduzir a influência do brilho do solo presente no NDVI, Huete (1988) propôs o índice de vegetação SAVI (*Soil-Adjusted Vegetation Index*), que usa um fator de ajuste de solo *L*:

$$SAVI = (L + 1)\frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red} + L} \ . \tag{3.2}$$

O valor de *L* normalmente é adotado como 1.

O índice EVI, ao contrário do NDVI, apresenta sensibilidade em áreas com maior quantidade de biomassa verde, além de minimizar os efeitos das influências atmosféricas e do solo (JIANG *et al*., 2008). Utilizando três comprimentos de onda, o cálculo do EVI é definido por:

$$EVI = G\frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + C_1 \times \rho_{red} - C_2 \times \rho_{blue} + L} \ , \tag{3.3}$$

onde G equivale a um fator de ganho, L corresponde ao fator de ajuste do solo, C_1 e C_2 são os coeficientes da resistência de aerossóis, que usam a reflectância na banda do azul (ρ_{blue}) para corrigir a influência dos aerossóis na banda do vermelho. No MODIS, os coeficientes geralmente utilizados para o cálculo do EVI são: $G = 2{,}5$, $L = 1$, $C_1 = 6$ e $C_2 = 7{,}5$ (JIANG *et al*., 2008).

Para obter um índice com comportamento similar ao EVI, com a prerrogativa de não utilizar a banda azul, Jiang *et al.* (2008) propuseram o EVI2 (*Enhanced Vegetation Index 2*), sugerindo não só uma melhora na sensibilidade sob

condições de elevada concentração de biomassa em relação ao SAVI, assim como uma minimização da influência do solo. O índice é calculado utilizando a reflectância nas bandas do infravermelho próximo e do vermelho:

$$EVI2 = 2{,}5\frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + 2{,}4 \times \rho_{red} + 1}\ . \tag{3.4}$$

3.1. Técnicas de detecção de mudanças

Devido à grande quantidade de dados de sensoriamento remoto disponíveis, técnicas de mineração de dados são utilizadas para facilitar a análise de padrões. Segundo Tan *et al*. (2001), na análise multitemporal os dados são usualmente compostos por uma sequência de *snapshots* (imagens obtidas em um determinado instante no tempo) que podem incluir variáveis com informações atmosféricas, terrestres e oceânicas, como ilustra a Figura 3.2.

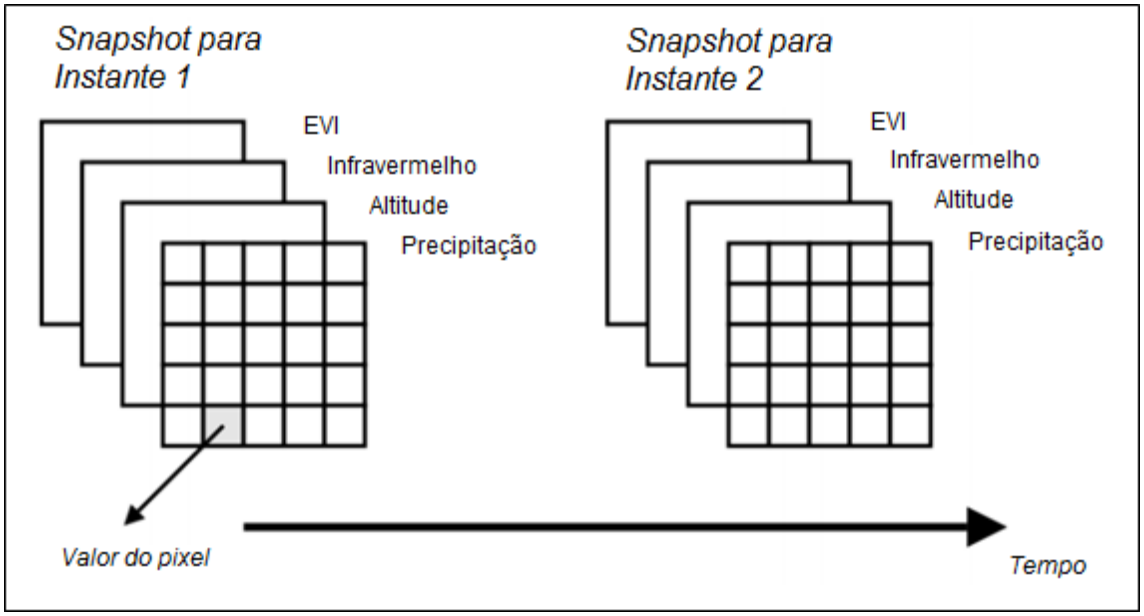


Figura 3.2. Visão simplificada do domínio do problema de análise temporal.
Fonte: Adaptada de Tan *et al*. (2001).

Além disso, a mudança observada depende significativamente do tipo de resolução temporal e deve ser compatível com os eventos analisados, que podem apresentar diferentes escalas espaço-temporais. Conforme mostra a Figura 3.3, muitos eventos utilizados na análise temporal são categorizados por intervalos de tempo heterogêneos.

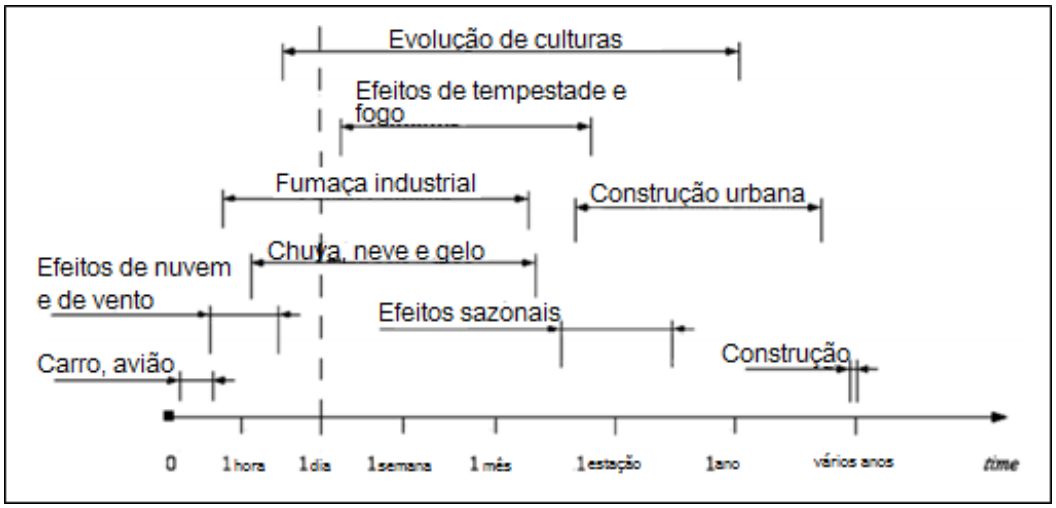


Figura 3.3. Intervalos de eventos no tempo.
Fonte: Adaptada de Heas e Datcu (2005).

É importante salientar que algumas condições devem ser satisfeitas antes de se implementar alguma técnica de detecção de mudanças. De acordo com Lu *et al*. (2004), é necessário que o registro e as calibrações atmosférica e radiométrica sejam precisas, além de, sempre que possível, fazer a seleção de imagens com a mesma resolução espacial e espectral. Além disso, é fundamental que fatores que ocasionem ruídos, como falhas nos dados e cobertura de nuvens, também sejam tratados.

Segundo Lu *et al.* (2004), as técnicas de detecção de mudanças podem ser sintetizadas em categorias, dentre elas: álgebra, transformação, interpretação visual e classificação.

3.1.1. Álgebra

Nesta categoria, os algoritmos utilizam operações algébricas para identificar mudanças, tais como diferença de imagens e a diferença entre índices de vegetação. Ferreira *et al*. (2013a), por exemplo, utiliza índices de vegetação extraídos de imagens *Landsat*-TM e MODIS para avaliar as propriedades biofísicas das áreas de pastagens no Cerrado brasileiro. Apesar de serem relativamente simples, os métodos algébricos não são capazes de fornecer matrizes com informações completas das mudanças, além da dificuldade na seleção de limiares para identificá-las (LU *et al*., 2004).

3.1.2. Transformação

Dentre os métodos na categoria de transformação, o PCA (*Principal Component Analysis*) é uma das aplicações mais comumente utilizadas. Com o objetivo de eliminar a redundância entre as bandas da imagem, este método obtém informações diferenciadas entre os componentes derivados da transformação. Estudos sobre expansão urbana (LI; YEH, 1998) e mortalidade de florestas (COLLINS; WOODCOCK, 1996) têm usado a técnica PCA para análise das mudanças. Porém, assim como os algoritmos algébricos, estas técnicas não fornecem matrizes com informações detalhadas e requerem a seleção de limiares para identificação das modificações. Além disso, a interpretação e rotulação das informações resultantes da transformação tornam-se difíceis (LU *et al.*, 2004).

O modelo linear de mistura espectral é outro método incluso nesta categoria e utilizado em várias aplicações na área de sensoriamento remoto. A seguir, conceitos e trabalhos que empregam este modelo são ilustrados.

3.1.2.1. Modelo linear de mistura espectral

O Modelo Linear de Mistura Espectral (MLME) parte da pressuposição que a resposta espectral de cada pixel da cena é uma mistura linear da resposta de vários alvos. A relação linear é utilizada para representar a mistura espectral

dos componentes dentro do elemento de resolução do sensor, de forma que a resposta de cada pixel, em qualquer banda espectral, pode ser definida como uma combinação linear das respostas de cada componente, desde que ele esteja presente no alvo imageado. Deste modo, por meio da extração das respostas individuais de cada alvo a partir de amostras de pixels puros desses alvos, também chamados de *endmembers*, pode-se modelar o peso com que cada alvo está contribuindo para o sinal de cada pixel e, assim realizar uma composição deste sinal em imagens-fração (ROBERTS *et al.*, 1998).

Segundo Teixeira (2004), alguns modelos de mistura tem sido propostos por Mascarenhas e Correia (1982), Shimabukuro (1987), Abrahão *et al.* (1990), Adams *et al.* (1990) e Pereira (1996). Alguns autores sugerem que tais misturas são compostas basicamente por três componentes: vegetação, solo e sombra (SHIMABUKURO *et al.*, 1997; AGUIAR, 1991). Segundo Shimabukuro e Smith (1991), e Shimabukuro (1987), estes três alvos são os elementos básicos de cenas florestadas, e o modelo de mistura espectral pode ser descrito por:

$$r_{i,j}(k) = a_{i,j} * veget + b_{i,j} * solo + c_{i,j} * sombra + \epsilon_{i,j}\ , \qquad (3.5)$$

onde *r* é o valor do pixel (*i*,*j*) na banda *k*, que varia de banda para banda e de pixel para pixel; $a_{i,j}$, $b_{i,j}$ e $c_{i,j}$ correspondem às proporções de vegetação, solo e sombra, respectivamente, em cada pixel. Os termos *veget*, *solo* e *sombra* equivalem às respostas espectrais de pixels puros, formados apenas pelos componentes de vegetação, solo e sombra, respectivamente. O valor de ϵ computa o erro na banda *k* e expressa a diferença entre a reflectância do pixel observado *r* e a reflectância do pixel computado a partir do modelo. A Figura 3.4 ilustra a decomposição de uma imagem *Landsat*-5 TM em imagens-fração de solo, sombra e vegetação.

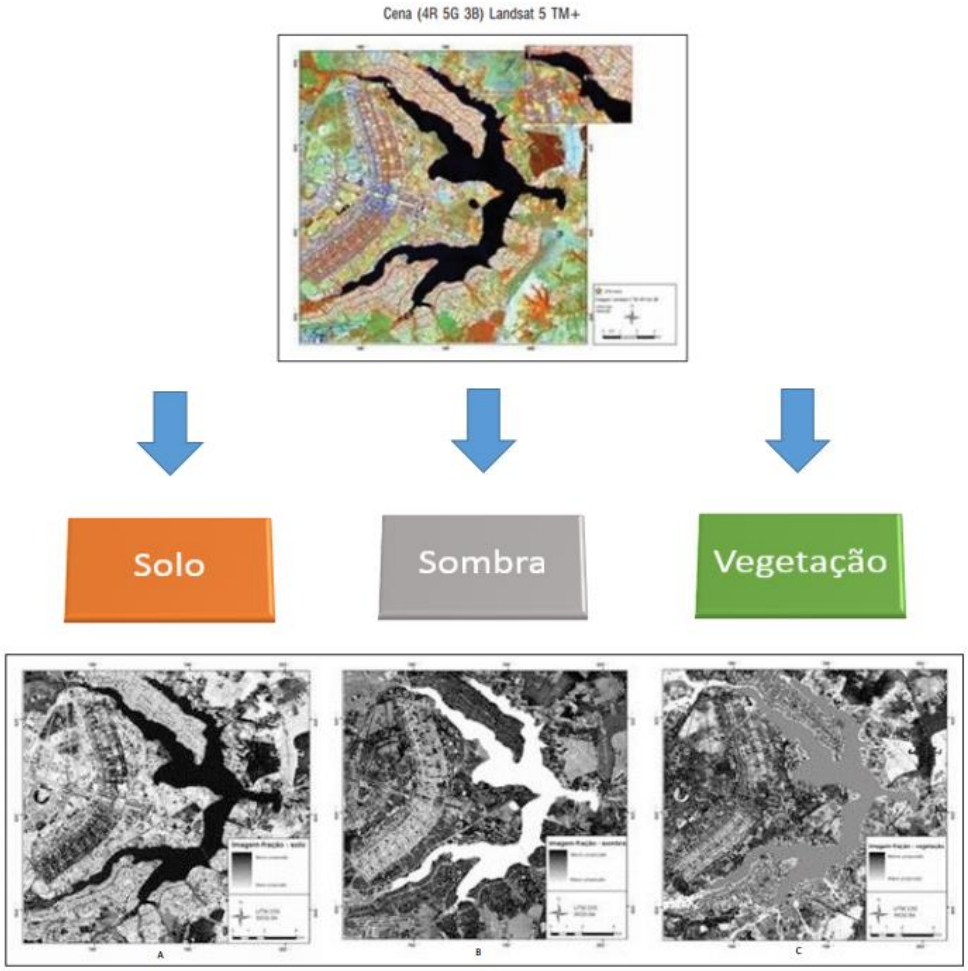


Figura 3.4. Exemplo da aplicação do MLME e suas imagens-fração resultantes.
Fonte: Adaptada de Bias *et al*. (2013).

As proporções são estimadas de maneira que a combinação das assinaturas espectrais dos componentes seja a melhor aproximação do valor do pixel observado. Os modelos baseiam-se no critério dos mínimos quadrados, para fazer a estimativa das proporções minimizando a soma dos quadrados dos erros $\epsilon_{i,j}$. Dentre os métodos utilizados, pode-se citar:

- Mínimos Quadrados com Restrições: método simples e rápido, é empregado quando o número de componentes é igual a três. A proporção de cada componente vegetação, solo e sombra deve estar

no intervalo de 0 e 1, e a soma das proporções dos três elementos deve ser igual a 1.

- Combinação entre Transformação de Principais Componentes e Mínimos Quadrados: reduzir o número de equações no sistema aplicando, primeiramente, a transformação de principais componentes para, em seguida, empregar o método de estimação de mínimos quadrados. Quando o número de componentes é maior do que três, esta técnica apresenta rapidez computacional em relação aos outros métodos.

O MLME vem sendo aplicado em vários estudos na área de sensoriamento remoto. Freitas e Cruz (2005) analisaram o modelo para a discriminação de classes de vegetação na Mata Atlântica. Freitas e Shimabukuro (2008), por exemplo, propuseram um método para identificar a ocorrência de desmatamento combinando o modelo linear de mistura com transformadas *wavelet*. Em relação ao Cerrado brasileiro, Mello *et al.* (2008) e Silva *et al.* (2010) avaliaram a viabilidade da aplicação do MLME na cobertura vegetal e na discriminação de fitofisionomias do bioma.

3.1.3. Interpretação visual

A categoria envolve a interpretação visual de imagens multitemporais através da composição de falsa cor, de maneira que características como textura, forma, tamanho e padrões das imagens são elementos que podem auxiliar na identificação das mudanças. Contudo, por meio de interpretação visual, é difícil gerar as trajetórias de mudanças detalhadas, aliado à grande demanda de tempo e dependência de especialistas para realizar a interpretação.

3.1.4. Classificação de imagens

Algoritmos de classificação de imagens são utilizados em várias aplicações de sensoriamento remoto. Segundo Lu *et al*. (2004), nos métodos de classificação de imagens, a qualidade e a quantidade de dados de amostra de treinamento

são essenciais para produzir resultados com boa acurácia. As vantagens do uso destes algoritmos no processo de análise de imagens estão na sua capacidade de criar uma matriz de informações de mudanças e na redução de impactos externos oriundos das diferenças ambientais e atmosféricas entre as imagens multitemporais. No entanto, o processo de seleção de amostras de treinamento numa classificação supervisionada para produzir mapas temáticos com acurácia satisfatória pode se tornar um problema complexo. A seguir, alguns métodos de classificação de imagens, baseados em técnicas de mineração de dados, são apresentados.

4. MINERAÇÃO DE DADOS

Com a digitalização das informações, a massiva quantidade de dados disponibilizados diariamente está cada vez maior. No entanto, a nossa habilidade de analisar e compreender um grande e crescente volume de dados encontra-se muito aquém se comparada a nossa capacidade de armazenamento destes. Com isso, surge a necessidade de extrair conhecimento potencialmente de alto nível a partir de dados de baixo nível (FAYYAD *et al.*, 1996).

A descoberta de conhecimento em um banco de dados (KDD – *Knowledge Discovery in Databases*) é todo o processo de descobrir conhecimento útil que não está explícito nos dados a serem trabalhados, sendo composto por várias etapas interligadas, conforme exposto na Figura 4.1. O KDD é um processo interativo e iterativo, com interconexões entre seus passos. A mineração de dados (*Data Mining*) representa um dos passos intermediários deste processo, e envolve a descoberta de modelos apropriados para um conjunto de dados e a determinação de padrões a partir destes (FAYYAD *et al.*, 1996).

Figura 4.1. Etapas do KDD. A mineração de dados corresponde a uma das etapas do processo da transformação de dados em conhecimento.

Fonte: Adaptada de Fayyad *et al.* (1996).

Segundo Lambin e Lindman (2006), a riqueza de dados de sensoriamento remoto disponíveis atualmente permite explorar questões sobre o sistema terrestre em escalas anteriormente infactíveis. Os autores relatam que a maioria das pesquisas em mudanças na cobertura do solo vinha sendo conduzidas pela disponibilidade de dados. Porém, com o aumento do número

de sensores e, consequentemente, do volume de informações, as pesquisas passaram a ser guiadas pela necessidade de processamento e análise dos dados, considerando que um melhor entendimento dos conjuntos de dados é essencial para a análise das mudanças da cobertura de solo e dos impactos dos processos do sistema terrestre. Portanto, o crescente montante de dados de sensoriamento remoto torna-se um caso apropriado para a aplicação de técnicas de mineração de dados (RUSHING *et al.*, 2005).

A classificação faz a predição e o mapeamento dos dados de um conjunto em uma ou mais classes categóricas (FAYYAD *et al.*, 1996; MILLER; HAN, 2009). A geração de modelos de classificação de dados é um processo composto de duas fases: aprendizado e teste. Na primeira etapa, aplica-se um algoritmo de classificação sobre um conjunto de dados de treinamento e, como resultado, tem-se a construção de um modelo de classificação, o classificador propriamente dito. O conjunto de treinamento, geralmente, equivale a um conjunto de amostras escolhidas aleatoriamente a partir da base de dados que se deseja analisar. Cada instância do conjunto de treinamento apresenta dois tipos de atributos: o atributo classe, o qual indica a classe que a instância pertence; e os atributos preditivos, cujos valores são avaliados a fim de descobrir o modo como eles relacionam-se com o atributo classe (HAN; KAMBER, 2006). Como é fornecida a classe a qual cada instância do conjunto de treinamento pertence, esta fase é conhecida como aprendizagem supervisionada. Esta se opõe à aprendizagem não supervisionada, na qual não se conhece a que classe pertence a coleção de dados, de modo que o número ou conjunto de classes a serem aprendidos pode também não ser antecipadamente conhecido.

Segundo Han e Kamber (2006), a fase de teste inicia-se após o classificador ser construído e seu objetivo é estimar e analisar a acurácia usando um conjunto de dados de teste. As amostras deste conjunto também são selecionadas de maneira aleatória, a partir da base de dados, e não podem ser utilizadas na construção do classificador. Ou seja, as amostras de teste devem

ser diferentes das amostras que foram escolhidas para gerar o conjunto de treinamento. A acurácia de um classificador em um conjunto de teste caracteriza a porcentagem de amostras que são corretamente classificadas por ele. Se a acurácia é considerada aceitável, pode-se utilizar o classificador para classificar novas amostras em que o rótulo de classe não é conhecido.

Antes de realizar a classificação e predição, é importante que os dados sejam pré-processados através, por exemplo, da redução de ruídos e tratamento de valores ausentes para um determinado atributo. Tais procedimentos podem auxiliar na redução da confusão durante a fase de aprendizagem. Além disso, quando métodos que envolvem medições de distâncias são utilizados na fase de aprendizado, os dados podem ser transformados através da normalização. Geralmente, este processo envolve ajustar para um conjunto de atributos todos os valores em uma escala, de forma que estes se encontrem em um intervalo específico, como de -1,0 a 1,0, ou de 0,0 a 1,0, podendo evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, possam influenciar o classificador de forma tendenciosa.

Algoritmos de mineração de dados são muito utilizados em aplicações de sensoriamento remoto. Por exemplo, classificação com Redes Neurais Artificiais (RNAs) tem sido usada para detectar mudanças em áreas urbanas (LIU; LATHROP, 2002) e em florestas (WOODCOCK *et al*., 2001; MAS *et al*., 2004). Costa (2009) faz a análise multitemporal por meio de cadeias difusas de Markov (*Fuzzy Markov Chain*, FCM). A partir de amostras de treinamento são gerados valores, chamados *fuzzy label vectors*, para um objeto em diferentes intervalos de tempo. Desta forma, a partir da matriz de transição baseada nestes valores, o método estima a classe a que cada novo objeto pertence no período de tempo seguinte, através de inferência *fuzzy*.

Bovolo *et al*. (2010) utilizam vetores de suporte para detecção de mudanças em imagens multitemporais e multiespectrais. Leite *et al*. (2011) utilizam um classificador baseado nas cadeias escondidas de Markov (*Hidden Markov*

Model, HMM) para identificar diferentes culturas agrícolas por meio da análise dos perfis temporais das características espectrais. Em outros estudos, como os de Smith (2010) e de Novack *et al.* (2011), os autores classificam a cobertura do solo usando a técnica de Florestas Aleatórias (do inglês, *Random Forests*).

Noma *et al*. (2013) comparam diversos métodos de classificação, como árvores de decisão, vizinhos mais próximos, Máquina de Vetores de Suporte (SVM) e Floresta de Caminhos Ótimos (OPF, *Optimun Path Forest*) para mapear áreas de vegetação. Os autores utilizam dados temporais correspondentes aos índices EVI em uma região do Mato Grosso com áreas de agricultura, floresta e pastagem. Neste caso, o método SVM apresentou os melhores resultados para o estudo em questão.

As técnicas de classificação baseadas em árvores de decisão, descritas na seção seguinte, têm sido utilizadas para mapeamento da vegetação (COLSTOUN *et al*., 2003), análise temporal de culturas agrícolas (KORTING, 2012; MELNYCHUK, 2012) e mapeamento de uso e cobertura do solo urbano (PINHO *et al.,* 2008; PINHO *et al*., 2012).

4.1. Classificação usando árvores de decisão

Em geral, uma árvore de decisão baseia-se em um conjunto de decisões aplicadas nos atributos disponíveis nos dados e são formadas, recursivamente, por meio da técnica de dividir e conquistar (HASTIE *et al.*, 2009). Numa árvore de decisão, cada nó interno caracteriza um teste em um atributo preditivo, uma ramificação partindo de um nó interno é um resultado para um teste e um nó folha representa um rótulo de classe.

A Figura 4.2 ilustra um exemplo típico de uma árvore de decisão, que determina se uma região é classificada como campo ou floresta. Amostras que satisfazem uma condição são atribuídas a uma ramificação da árvore. Este processo é repetido de forma recursiva para cada ramo da árvore, e quando

algum ramo contém apenas informações de uma única classe, este se torna uma folha. Após construída a árvore, um nó folha corresponde a um valor categórico esperado para todas as instâncias descritas no caminho entre a raiz e a folha (THEODORIDIS; KOUTROUMBAS, 2008).

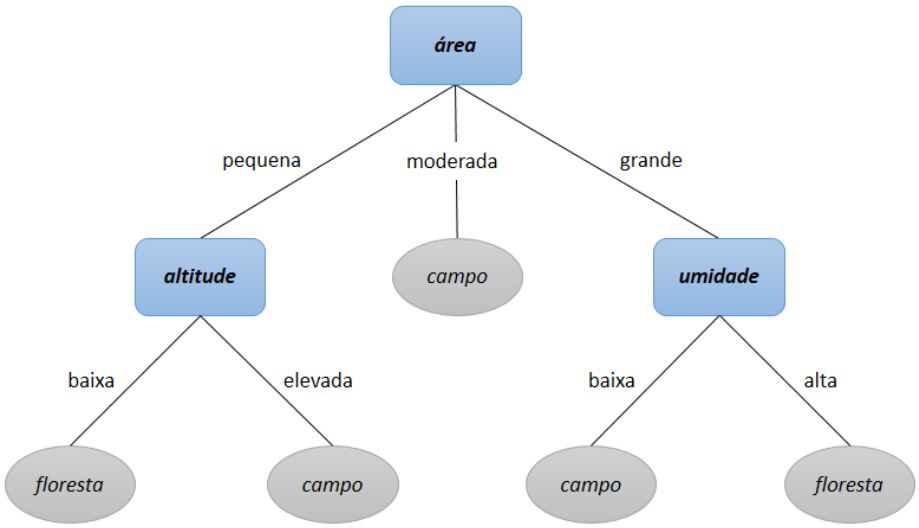


Figura 4.2. Exemplo de árvore de decisão indicando se uma região é uma provável floresta ou um campo.

É importante ressaltar que alguns algoritmos geram apenas árvores de decisão binárias, enquanto outros podem produzir modelos não binários (HAN; KAMBER, 2006). Além disso, uma árvore de decisão pode lidar com dados de alta dimensão e sua construção não exige configuração de parâmetros e nem domínio do conhecimento, sendo apropriada para a descoberta exploratória de conhecimento (HAN; KAMBER, 2006). Segundo McCauley e Goetz (2004), o processo de indução da árvore de decisão é claro, rápido e intuitivo, tornando o modelo de classificação fácil de ser interpretado.

Uma questão chave para a geração de uma árvore de decisão consiste na estratégia para a escolha da posição dos atributos preditivos na árvore. Para isso, os atributos preditivos que serão utilizados em cada nó são determinados a partir de um critério de seleção. Geralmente, divide-se os dados de um nó pai de forma a minimizar o grau de impureza dos nós filhos, de modo que o atributo com o maior valor para o critério de seleção seja escolhido como

atributo de divisão para o conjunto de amostras. O nó criado para a partição é rotulado com o critério de separação, e as ramificações das árvores crescem a partir do resultado desta condição, particionando o conjunto de amostras por meio desta (HAN; KAMBER, 2006).

Um dos métodos comumente utilizados equivale ao ganho de informação, que utiliza a entropia como medida, a qual caracteriza a impureza de uma coleção arbitrária de amostras. Seja *D* um conjunto de treinamento e suponha que o atributo classe contenha c valores distintos, correspondente à quantidade de classes C_i no conjunto *D*, com $i = 1, \dots, c$. Baseando-se nestes dados, a entropia pode ser definida como:

$$Entropia(D) = -\sum_{i=1}^{c} p_i \log_2 p_i \,, \tag{4.1}$$

onde p_i é a probabilidade que o conjunto arbitrário de amostras pertence à uma classe C_i (WITTEN; FRANK, 2000; HAN; KAMBER, 2006).

Suponha agora que se queira particionar as amostras em *D* através de um atributo preditivo *A* que contém *k* valores distintos, $\{a_1, a_2, \dots, a_k\}$. O atributo *A* pode ser utilizado para dividir *D* em *k* partições, $\{D_1, D_2, \dots, D_k\}$, onde D_j representa as amostras em *D* que tenham o valor a_j. Sendo |.| o operador de módulo, o valor esperado da entropia em *D,* se particionado pelo atributo *A,* pode ser definido:

$$Entropia_A(D) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} \times Entropia(D_j) \tag{4.2}$$

Por conseguinte, partindo das equações 4.1 e 4.2, o ganho de informação de um atributo preditivo *A* corresponde a redução esperada na entropia pelo particionamento das amostras a partir de *A*, expresso por:

$$Ganho(A) = Entropia(D) - Entropia_A(D) \quad (4.3)$$

O atributo *A* que tiver o maior valor de ganho de informação é escolhido como o atributo de particionamento no nó da árvore. Para ilustrar o processo de seleção de atributos para os nós das árvores, considere, por exemplo o conjunto de treinamento da Tabela 4.1. A tabela representa um conjunto de treinamento *D* com 14 amostras, com quatro atributos preditivos para cada amostra. O atributo classe *vegetação* tem dois valores distintos (*campo* e *floresta*), ou seja, *c* = 2. Têm-se, então, nove amostras da classe *campo* e cinco da classe *floresta*.

Tabela 4.1 – Amostras de treinamento para uma base de dados de cobertura do solo.
Fonte: Adaptada de Quilan (1986)

N.	***Atributos preditivos***				***Atributo classe***
	área	***altitude***	***umidade***	***vento***	***vegetação***
01	pequena	elevada	baixa	fraco	floresta
02	pequena	elevada	baixa	forte	floresta
03	moderada	elevada	baixa	fraco	campo
04	grande	média	baixa	fraco	campo
05	grande	baixa	alta	fraco	campo
06	grande	baixa	alta	forte	floresta
07	moderada	baixa	alta	forte	campo
08	pequena	média	baixa	fraco	floresta
09	pequena	baixa	alta	fraco	campo
10	grande	média	alta	fraco	campo
11	pequena	média	alta	forte	campo
12	moderada	média	baixa	forte	campo
13	moderada	elevada	alta	fraco	campo
14	grande	média	baixa	forte	floresta

Seja a classe C_1 equivalente a *campo* e a classe C_2 equivalente a *floresta*, então, pode-se calcular a entropia para o conjunto *D* a partir da Equação 4.1:

$$Entropia(D) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0{,}940.$$

A seguir, calcula-se, para cada um dos atributos preditivos, o valor de entropia esperado caso este seja escolhido. Para o atributo *área*, por exemplo, tem-se:

$$\begin{aligned} Entropia_{área}(D) = {} & \frac{5}{14}\times\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ & + \frac{4}{14}\times\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\ & + \frac{5}{14}\times\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ = {} & 0{,}694. \end{aligned}$$

Por conseguinte, o valor do ganho de informação para o atributo *área* é:

$$Ganho(área) = Entropia(D) - Entropia_{área}(D) = 0{,}940 - 0{,}694 = 0{,}246.$$

De forma análoga, calculam-se os ganhos de informação para os outros atributos preditivos, obtendo-se $Ganho(altitude) = 0{,}029$, $Ganho(umidade) = 0{,}151$ e $Ganho(vento) = 0{,}048$. Como o atributo *área* alcançou o maior valor de ganho de informação, ele é utilizado como atributo de decisão no nó da árvore, criando uma ramificação na árvore de decisão.

Por outro lado, vale ressaltar que o ganho de informação favorece atributos com muitos valores possíveis. Por exemplo, ao utilizar um atributo que não tenha relevância, como um identificador único, seria criado um nó para cada valor possível e o número de nós seria igual ao número de identificadores. Cada um dos nós teria somente uma amostra e o valor de entropia seria mínimo já que em cada nó todas as amostras pertencem à mesma classe. Essa divisão geraria um ganho máximo, embora sem utilidade.

Para contornar esta situação, Quilan (1993) propôs o classificador C4.5, no qual se calcula a razão de ganho (do inglês, *Gain Ratio*), uma extensão do ganho de informação que favorece atributos em que a entropia apresenta valor baixo, superando o ganho de informação em termos de acurácia e de complexidade das árvores geradas. A razão de ganho equivale ao ganho de informação relativo (ponderado) como critério de avaliação e é definido por:

$$Razão\ de\ Ganho(A) = \frac{Ganho(A)}{Entropia\ de\ Particionamento_A(D)}, \tag{4.4}$$

onde a entropia de particionamento, do inglês *Split Info*, comporta-se de forma análoga à $Entropia_A(D)$, penalizando amostras com um número grande de partições, que é equivalente a:

$$Entropia\ de\ Particionamento_A(D) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{4.5}$$

No cálculo de razão de ganho para o atributo *altitude* da Tabela 4.1, utilizando a Equação 4.5, encontrar-se-á:

$$Entropia\ de\ Particionamento_{altitude}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right)$$

$$= 1{,}557.$$

Pode-se, então, obter:

$$Razão\ de\ Ganho(altitude) = \frac{Ganho(altitude)}{Entropia\ de\ Particionamento_A(altitude)} = \frac{0{,}029}{1{,}557}$$

$$= 0{,}019.$$

Calculando a razão de ganho para os atributos *área, umidade* e *vento*, os valores obtidos são 0,157, 0,152 e 0,049, respectivamente. O atributo *área* contém o maior valor de razão de ganho e, consequentemente, seria o atributo selecionado para compor o nó.

4.2. Classificação usando florestas aleatórias

Existem algumas técnicas que combinam modelos de aprendizado com o objetivo de melhorar a acurácia da classificação. Tipicamente, algoritmos de combinação de modelos geram um conjunto amplo de classificações, a partir dos dados, utilizando um algoritmo de aprendizagem base (FRIELD *et al.*, 1999; BREIMAN, 2001; RODRIGUEZ-GALIANO *et al.*, 2012). Dentre estes, há a técnica de *Random Forests*. Proposta por Breiman (2001), Florestas Aleatórias é um algoritmo de aprendizagem que combina um grupo de classificadores baseados em árvores de decisão.

A técnica consiste em um conjunto de classificadores estruturados em árvores $\{h(X, \Theta_k), k = 1, \ldots.\}$, onde $\{\Theta_k\}$ são vetores amostrados de forma independente e distribuídos identicamente para todas as árvores da floresta. Os vetores são gerados aleatoriamente por meio de uma distribuição de probabilidade fixa sobre o conjunto de treinamento (BREIMAN, 2001). Depois da geração de todas as árvores de decisão, cada uma lança um voto para uma classe na entrada *X*. Com isso, a classe mais votada é a selecionada na predição do classificador (GHIMIRE *et al.*, 2010). A Figura 4.3 ilustra o processo de predição para uma amostra.

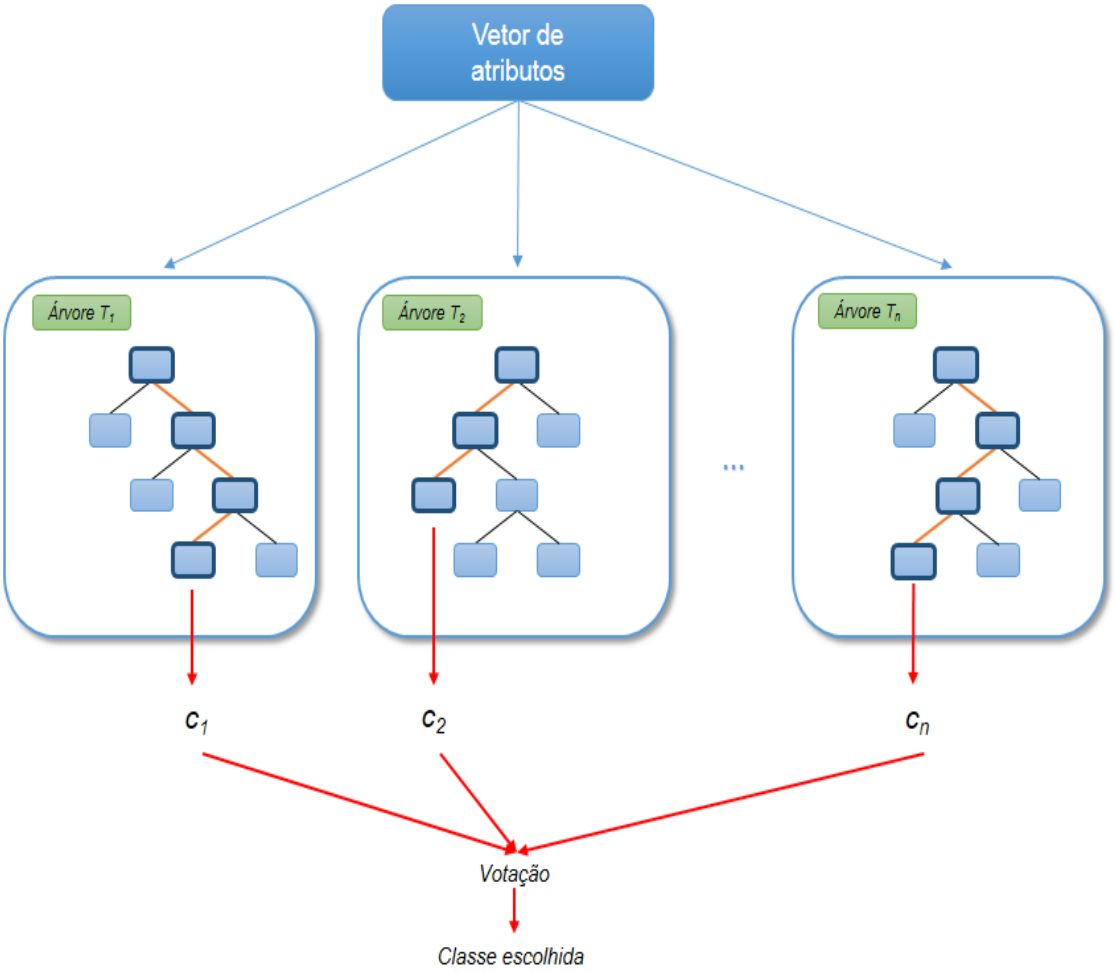


Figura 4.3. Exemplo do processo de seleção de uma classe a partir da classificação em floresta aleatória.

Algoritmos de combinação de modelos podem beneficiar o mapeamento de cenários complexos, uma vez que a natureza iterativa destes métodos produz classificações múltiplas e garante uma melhora na acurácia, visto que o erro de uma única classificação é sobreposto pela combinação de múltiplas classificações (FRIELD *et al.*, 1999; SESNIE *et al.*, 2010; RODRIGUEZ-GALIANO *et al.*, 2012).

As florestas aleatórias herdam a natureza determinística da classificação por árvores de decisão e, como diversas árvores são geradas, garante-se que os padrões nos dados sejam generalizados e possam ser aplicados em amostras

em que a classe não é conhecida (BREIMAN, 2001). Além disso, apenas dois parâmetros precisam ser ajustados: a quantidade de árvores geradas e a quantidade dos atributos preditivos utilizados (GHIIMIRE *et al.,* 2012; RODRIGUEZ-GALIANO *et al.*, 2012).

Alguns trabalhos na área de sensoriamento remoto usam a técnica de florestas aleatórias no processo de classificação. Rodriguez-Galiano *et al.* (2012), por exemplo, utilizam as florestas aleatórias na classificação da cobertura do solo, com acurácia de 92% para a área de estudo em uma região da Espanha. Pal (2005) e Sensie *et al.* (2010) relatam que tais modelos apresentam acurácia similar aos algoritmos de aprendizagem mais complexos como o SVM, por exemplo.

4.3. Classificação usando SVM

Uma outra abordagem que vem recebendo crescente atenção é aquela baseada em máquinas de vetores de suporte (*Support Vector Machines*, SVM) (BOSER *et al*., 1992; CORTES; VAPNIK, 1995). Embasado na Teoria da Aprendizagem Estatística (VAPNIK, 1995), o método tem sido eficiente em classificações com dados de alta dimensionalidade, muitas vezes com resultados superiores aos de outros algoritmos de aprendizado, como as redes neurais (SUNG; MUKKAMALA, 2003; DING; DUBCHAK, 2001).

Método de aprendizagem supervisionada capaz de classificar dados lineares e não lineares, o SVM utiliza, de forma geral, um mapeamento não linear para transformar os dados do conjunto de treinamento original em uma nova dimensão. Dentro desta nova dimensão, o algoritmo faz uma busca pelo limite de decisão linear ótimo para separar duas classes: o hiperplano cuja margem de separação seja máxima (HAN; KAMBER, 2006).

Embora o tempo da fase de treinamento dos algoritmos de SVM possa ser considerado moderado, a precisão deles é alta devido a sua capacidade de modelar limites de decisão não lineares complexos, apresentando uma boa

generalização. Como é um classificador não-paramétrico, o SVM tem como característica a independência de modelos de distribuição estatística (MELGANI; BRUZZONE, 2004).

O caso mais simples que o algoritmo lida é o de duas classes, em que os dados são linearmente separáveis. Suponha um conjunto de treinamento D com *n* amostras, $D = \{(x_i, y_i),\ i = 1, \ldots, n\}$, onde a amostra x_i está associada ao rótulo de classe y_i. Se utilizado o conjunto de treinamento da Tabela 4.1, por exemplo, cada y_i assumiria um de dois valores, -1 ou +1, correspondentes aos valores de exemplo, *campo* e *floresta*, respectivamente. Considerando um exemplo baseado em dois atributos de entrada, x_1 e x_2, como exposto na Figura 4.4, percebe-se que os dados são linearmente separáveis, uma vez que uma reta pode ser desenhada separando as amostras da classe -1 (círculos em verde) das amostras da classe $+1$ (círculos em laranja).

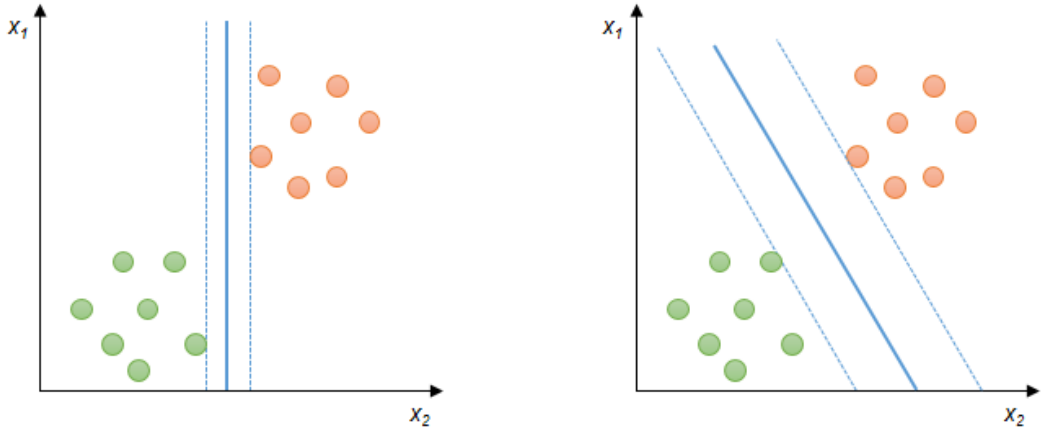


Figura 4.4. Exemplos de retas capazes de separar as classes -1 e +1. Há um número infinito de retas que separam as duas classes.

Como existem infinitas retas que podem separar essas classes, o problema, então, reside em encontrar aquela que mais se adequa no processo de classificação. Generalizando, procura-se encontrar a superfície de decisão entre as classes baseada no comportamento geométrico das amostras no espaço de atributos, ou seja, busca-se o hiperplano que minimize o erro de classificação. Para isso, o SVM pesquisa o hiperplano de máxima margem de

separação, onde a margem corresponde à distância mais curta do hiperplano para as amostras de treinamento mais próximas de cada classe (HAN; KAMBER, 2006). O hiperplano de separação pode ser definido como:

$$f(x) = \langle w, x \rangle + b = 0, \tag{4.6}$$

onde *w* é o vetor ortogonal ao hiperplano de separação, e *b* é um escalar tal que $|b|/\|w\|$ corresponde à distância do hiperplano à origem do espaço de atributos, onde $\|.\|$ representa a norma vetorial e $\langle .,. \rangle$ corresponde ao produto interno. As amostras localizadas nos limites da margem de separação (ou seja, quando $f(x) = -1$ ou $f(x) = +1$) são as responsáveis por determinar o hiperplano, e são chamadas de vetores de suporte. A Figura 4.5 ilustra a escolha de um hiperplano de seleção ótimo, com os vetores de suportes contornados com uma borda mais escura.

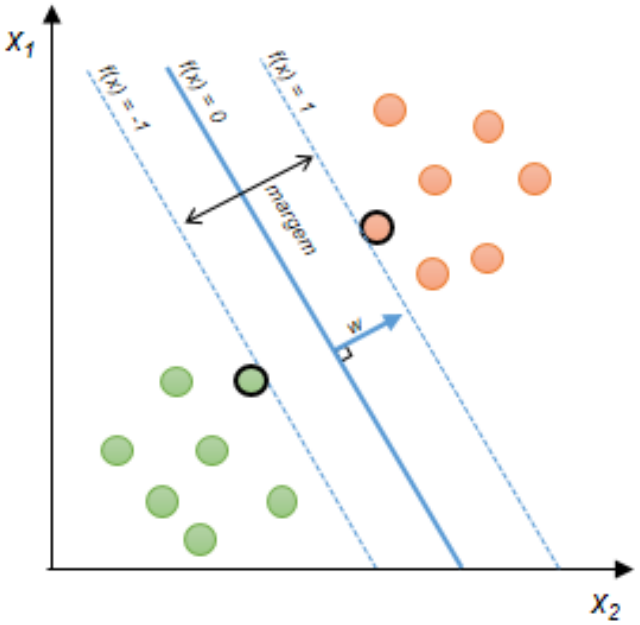


Figura 4.5. Representação do hiperplano ótimo e dos vetores de suporte.

A Equação 4.6 separa o espaço *x* em duas regiões, de maneira que uma função sinal, *g(x)*, pode ser aplicada na obtenção das classificações, como segue:

$$g(x) = \begin{cases} +1, se\ \langle w, x\rangle + b > 0 \\ -1, se\ \langle w, x\rangle + b < 0 \end{cases}, \tag{4.7}$$

Com isso, o hiperplano de margem máxima de separação deve ser definido de forma que as amostras mais próximas a ele satisfaçam a condição $|\langle w, x_i\rangle + b| = 1$, implicando nas inequações:

$$f(x_i) = \begin{cases} \langle w, x_i\rangle + b \geq +1\ se\ y_i = +1 \\ \langle w, x_i\rangle + b \leq -1\ se\ y_i = -1 \end{cases}, \tag{4.8}$$

Estas, então, podem ser combinadas em:

$$y_i(\langle w, x_i\rangle + b) \geq 1, i = \{1,2, \dots, n\}\ . \tag{4.9}$$

A partir da restrição imposta pela Equação 4.9, pode-se obter o valor da margem de separação, que tem largura $2/\|w\|$. Por conseguinte, infere-se que a minimização de $\|w\|$ leva à maximização da margem. Então, considerando as relações supracitadas, o algoritmo SVM encontra o hiperplano através da determinação dos parâmetros *w* e *b*, que pode ser modelado pelo seguinte problema de otimização (THEODORIDIS; KOUTROMBAS, 2008):

$$\begin{aligned} &\min_{w,b}\left(\frac{1}{2}\langle w, w\rangle\right), \\ &\text{sujeito a: } y_i(\langle w, x_i\rangle + b) \geq 1, i = \{1,2, \dots, n\}. \end{aligned} \tag{4.10}$$

O problema de otimização que define o hiperplano de separação ótima apresenta função objetivo convexa e pode ser solucionada com a introdução de uma função langrangiana, que adiciona as restrições à função objetivo,

associadas a parâmetros denominados multiplicadores de Langrange, α_i. Uma vez que os vetores de suportes e o hiperplano de margem máxima de separação são encontrados, constrói-se o classificador: a máquina de vetores de suporte, que pode ser utilizada para classificar dados linearmente separáveis.

Baseado na formulação langrangiana citada, o hiperplano de margem máxima de separação pode ser reescrito como um limite de decisão para ser utilizado como um teste de classificação (HAN; KAMBER, 2006):

$$d(x^T) = \sum_{i=1}^{l} y_i \alpha_i \langle x_i, x^T \rangle + b_0 \tag{4.11}$$

onde y_i é a classe associado ao vetor de suporte x_i; x^T é a amostra de teste que será classificada, α_i e b_0 são parâmetros que são determinados automaticamente pela otimização; e l é o número de vetores de suporte.

Para fazer a predição, uma amostra de teste x^T é inserida como parâmetro em 4.11 e, em seguida, observa-se o sinal de $d(x^T)$, como resultado da equação. Se for positivo, o SVM prediz que a amostra x^T pertence à classe +1. Caso contrário, a amostra é rotulada com a classe -1.

Contudo, aplicações em situações reais cujos dados sejam linearmente separáveis são bastante difíceis de ser encontradas. Para contornar este problema, o algoritmo SVM linear pode ser estendido para classificação de dados não linearmente separáveis, sendo capaz de descobrir limites de decisão não lineares no conjunto de amostra de treinamento. Para isso, faz-se a transformação dos dados de treinamento originais para um novo espaço dimensional através do mapeamento não linear. Com os dados transformados, o algoritmo procura por um hiperplano linear no novo espaço de atributos (HAN; KAMBER, 2006). A Figura 4.6 ilustra o processo de transformação dos dados de entrada em uma nova dimensão.

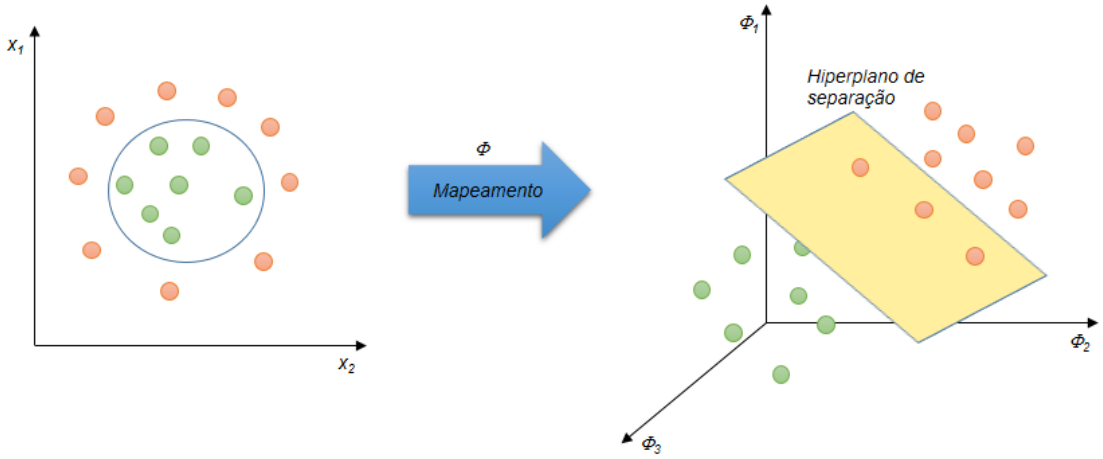


Figura 4.6. Exemplo de mapeamento para um espaço de maior dimensão, fornecendo uma separação linear.

A determinação dos parâmetros do hiperplano de separação na nova dimensão corresponde outra vez a um problema de otimização quadrática langrangiana. No entanto, quanto maior a dimensionalidade, maior será o custo computacional envolvido. Observando novamente a Equação 4.11, pode-se perceber que é necessário, para predizer uma amostra de teste x^T, calcular o produto de x^T com cada um dos vetores de suporte, o que pode tornar a tarefa bastante custosa.

No novo espaço dimensional, as amostras de treinamento aparecem na forma de produtos internos, $\langle \Phi(x_i), \Phi(x_j) \rangle$, onde $\Phi(x)$ equivale à função de mapeamento não linear nos dados de treinamento iniciais. Para resolver esta situação, pode-se substituir o produto interno presente no problema de otimização por uma função *Kernel* K, tal que $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ (THEODORIDIS; KOUTROMBAS, 2008). Desta forma, todos os cálculos são realizados nas amostras de treinamento iniciais, reduzindo potencialmente o custo computacional.

Alguns dos *kernels* mais empregados em aplicações gerais são as funções de base radial (*Radial Basis Functions,* RBF) e as polinomiais, descritos a seguir.

- Linear: $K(x_i, x_j) = \langle x_i, x_j \rangle$.
- Polinomial: $K(x_i, x_j) = (\langle x_i, x_j \rangle + 1)^d$.
- RBF: $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\gamma^2}}$.

Vale ressaltar que $d \in \mathbb{N}^*$ e $\gamma \in \mathbb{R}_+^*$são parâmetros que precisam de ajustes no processo de classificação.

Com a seleção do *kernel*, pode-se em seguida realizar o processo de decisão de forma similar ao processo de predição de dados linearmente separáveis. Neste caso, porém, há a introdução do parâmetro *C*, que deve ser ajustado e é responsável pela regularização associada à quantidade de amostras classificadas erroneamente. Geralmente, tanto o parâmetro *C* quanto o parâmetro γ do *kernel* RBF são obtidos empiricamente, através de pesquisa exaustiva por meio de validação cruzada.

Como vantagem, a complexidade do classificador é definida pelo número de vetores de suporte em vez da dimensionalidade dos dados. Além disso, se outras amostras de treinamento forem removidas e a fase de treinamento for repetida, e estas não sejam vetores de suporte, o mesmo hiperplano de separação será encontrado. Desta forma, o número de vetores de suporte pode ser usado para calcular o limite superior da taxa de erro esperado no classificador, independentemente da dimensão dos dados (HAN; KAMBER, 2006).

Embora o SVM seja, originalmente, esquematizado para uma classificação binária, é possível compor várias SVMs para tratar problemas com mais de duas classes. As principais abordagens usam como base a decomposição de um problema multiclasse em subproblemas binários e, em seguida, faz-se a reconstrução do problema por meio da combinação dos resultados destes subproblemas. Dois métodos se destacam: “Um-Contra-UM” (*One-Against-*

One) e “Um-Contra-Todos” (*One-Against-All*) (WEBB, 2002; THEODORIDIS; KOUTROMBAS, 2008).

Em problemas que usam a estratégia “Um-Contra-Todos” são aplicadas *c* SVMs, onde *c* é o número de classes. Cada classificador faz a separação de uma classe das demais, de forma que as amostras de uma determinada classe são associadas à classe +1, enquanto que todas as demais recebem indicadores de classe -1. No processo de decisão, usa-se o esquema de votação por maioria, onde cada um dos classificadores anuncia um rótulo de classe como resultado. A amostra de teste é predita com a classe que receber a maioria dos votos. Por outro lado, na estratégia “Um-Contra -Um” comparam-se classes duas as duas. São necessários $c(c-1)/2$ classificadores, cada um responsável pela separação das amostras entre um determinado par de classes. Analogamente, a classe selecionada é aquela mais votada no processo de predição do método.

5. METODOLOGIA

Neste capítulo é descrita a metodologia proposta para mapeamento de áreas de Pastagem Cultivada e Campo Nativo (Campo Limpo, Sujo e Rupestre).

Para validar a metodologia, foi escolhida como área de estudo uma região localizada no sudoeste do estado de Minas Gerais. A área faz parte do complexo do Cerrado, e apresenta regiões naturais e antrópicas, com áreas de pastagem cultivada. Nesta região, ilustrada na Figura 5.1, está presente o Parque Nacional da Serra da Canastra e regiões vizinhas, onde são encontradas formações naturais de Campo Limpo, Campo Sujo e Campo Rupestre (IBGE, 2012). Esta região foi escolhida por apresentar os alvos de interesse como as áreas campestres nativas e pastagens cultivadas.

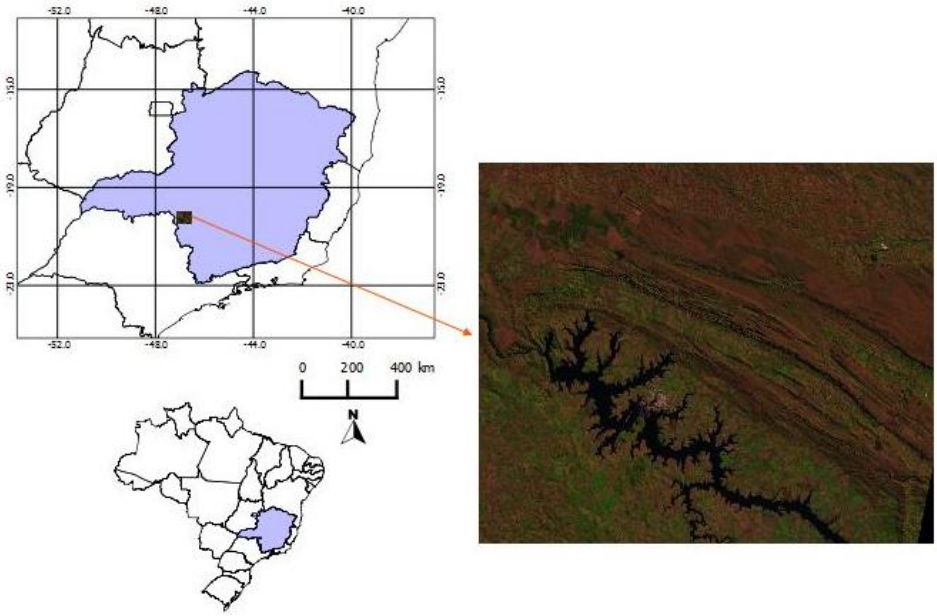


Figura 5.1 Área de estudo do trabalho, localizada no sudoeste do estado de Minas Gerais. Recorte de uma imagem *Landsat* TM-5 (R5B4G3) da região.

As etapas de processamento e análise dos dados propostas neste trabalho estão esquematizadas na Figura 5.2. A seguir, apresenta-se uma descrição

detalhada dos tipos de dados de entrada e as etapas de processamento dos dados que foram usados para o desenvolvimento desta pesquisa.

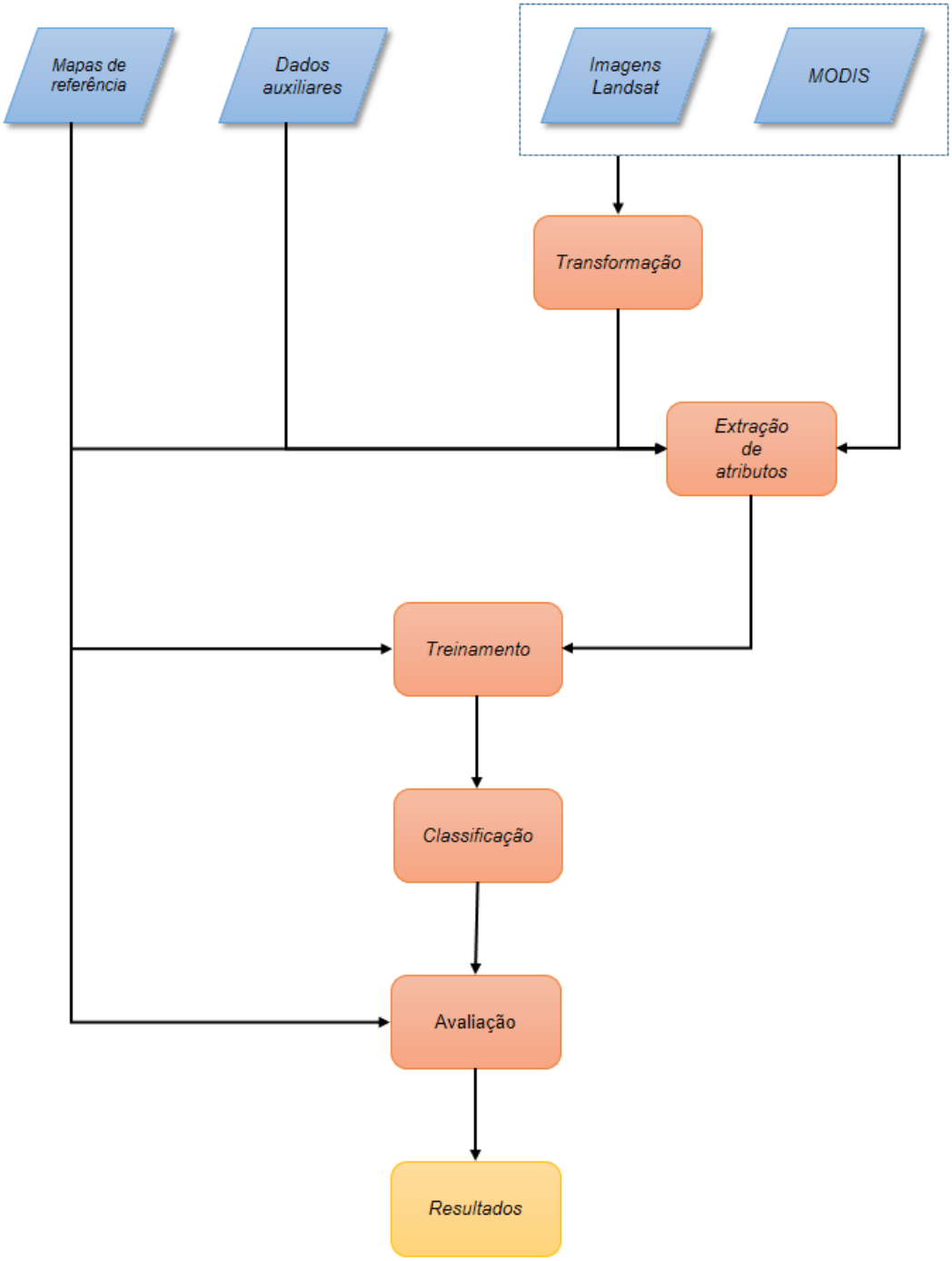


Figura 5.2 Etapas de processamento e análise dos dados.

5.1. Mapas de referência

Para a obtenção das formações campestres nativas (Campo Limpo, Sujo e Rupestre), o mapa de cobertura vegetal do Cerrado fornecido pelo Inventário Florestal de Minas Gerais do ano de 2009 foi utilizado como referência para treinamento e validação da classificação. O mapa é o resultado do esforço conjunto de diversas instituições, dentre as quais a Universidade Federal de Lavras apresenta papel de destaque na produção dos mapeamentos obtidos (SCOLFORO *et al.*, 2008). A Figura 5.3 exibe o mapa de cobertura da área de estudo.

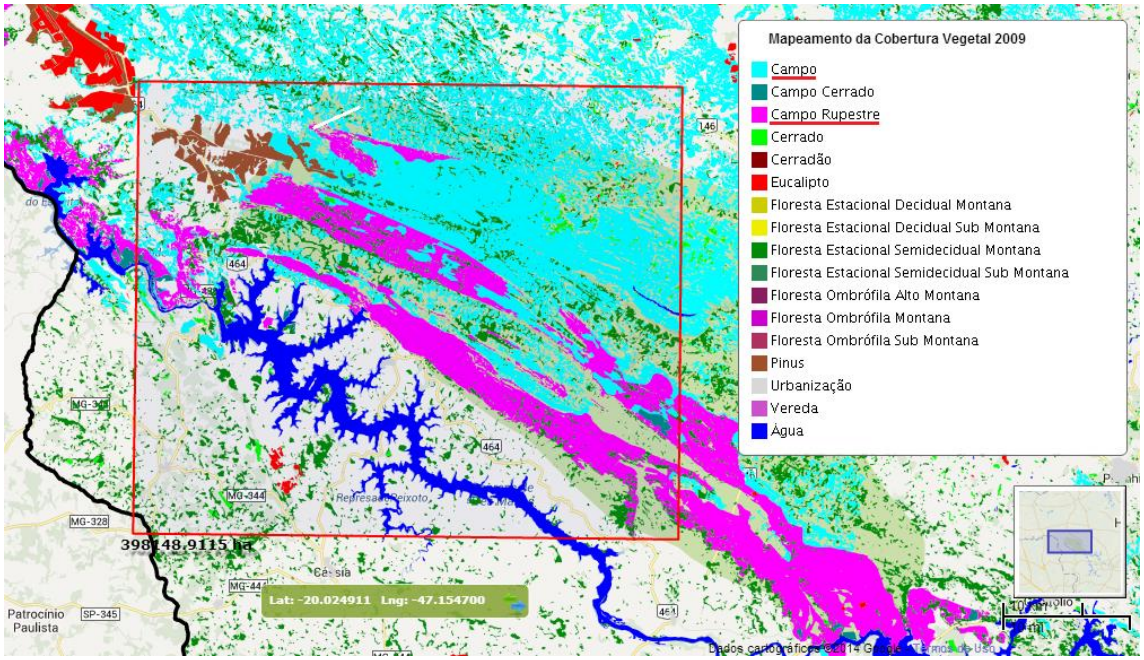


Figura 5.3 Mapa de referência utilizado na área de estudo para obtenção das regiões de formações campestres: Campo (cor ciano) e Campo Rupestre (cor rosa).

É importante ressaltar que o mapa de cobertura vegetal citado contém duas classes que representam as formações campestres do Cerrado: a classe Campo, que combina as fitofisionomias de Campo Limpo e Campo Sujo; e a classe Campo Rupestre, definida pela fitofisionomia homônima.

Em relação às pastagens cultivadas do Cerrado, o Ministério do Meio Ambiente (MMA) oferece o mapa, relativo ao ano de 2006, que distingue a cobertura antrópica da natural no bioma e faz a discriminação das áreas antrópicas do Cerrado, como mostra a Figura 5.4. O mapeamento subdivide as áreas antrópicas em quatro classes: agricultura, pastagem cultivada, região de florestamento/reflorestamento e área degradada por mineração. Dessa forma, o mapa do MMA foi também usado como referência para obtenção de áreas de pastagem cultivada.

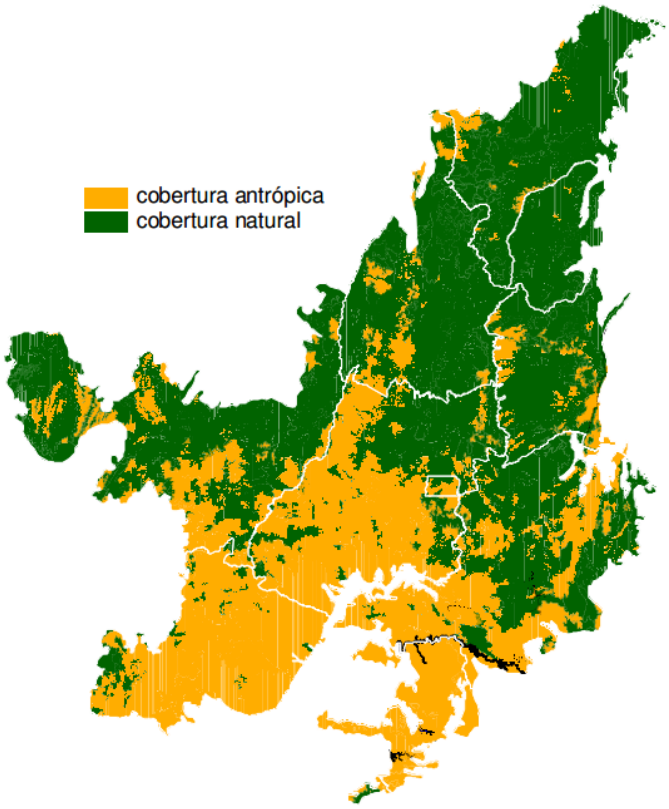


Figura 5.4 Regiões do Cerrado mapeadas pelo MMA como áreas naturais (em verde) e antrópicas (em amarelo).

Ambos os mapas de referência foram integrados em um único conjunto de regiões (ou polígonos) utilizando as ferramentas de álgebra de mapas disponíveis no *software* TerraView (INPE, 2013). Primeiramente, todos os polígonos do mapa de cobertura do Inventário de Minas Gerais

correspondentes às classes de formações vegetais, exceto Campo e Campo Rupestre, foram associados à classe Outros, considerando agora apenas estas três classes. Já no mapa do MMA apenas os polígonos atribuídos à classe Pastagem Cultivada foram considerados. Foi realizada a operação de diferença entre os dois conjuntos de polígonos seguida da operação de soma, com o intuito de eliminar polígonos com sobreposição e, consequentemente, regiões que apresentassem conflitos de classe.

Com isso, criou-se um arquivo *shapefile* com polígonos associados a uma das quatro seguintes classes: Pastagem Cultivada, Campo (Campo Limpo/Campo Sujo), Campo Rupestre e Outros. Estes polígonos são usados na etapas de extração de atributos e treinamento, e também como referência para avaliar os resultados da classificação.

5.2. Dados auxiliares

Segundo Ribeiro e Walter (2008), áreas de Campo Rupestre são encontradas geralmente em regiões de altitude elevada, superiores a 900 metros, ocasionalmente a partir de 700 metros. Por conseguinte, foram coletados dados com informações de topografia da área de estudo para auxiliar no processo de discriminação das classes. As informações foram obtidas a partir de dados TOPODATA (VALERIANO, 2005) da região com resolução espacial de 30 m, por meio da coleta da resposta dos valores de altitude e declividade fornecidos por estes dados. A Figura 5.5 ilustra as imagens que caracterizam os valores de altitude e declividade para a região de interesse.

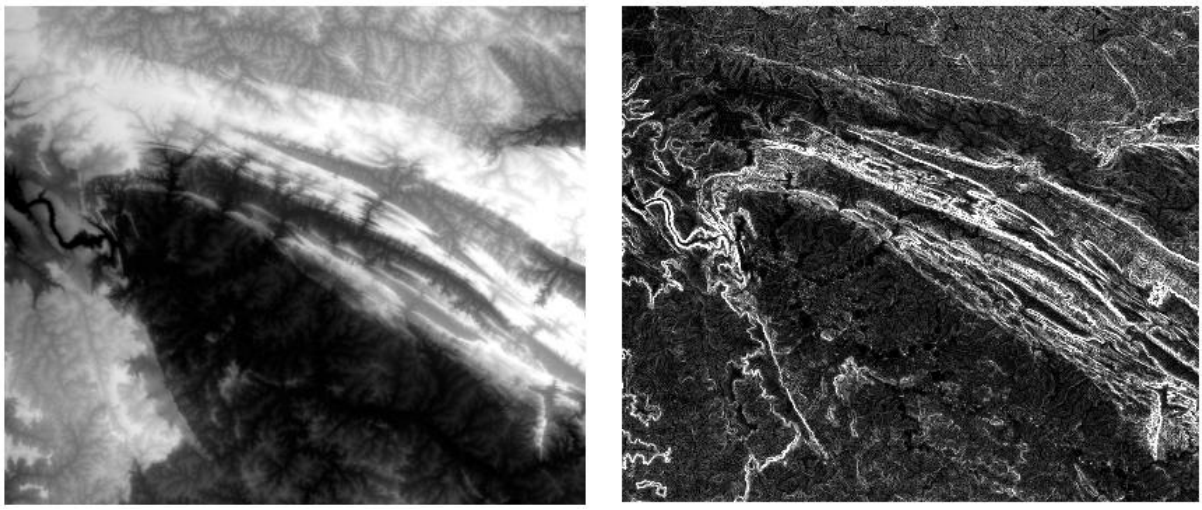


Figura 5.5 Imagens TOPODATA relativas à altitude (esquerda) e à declividade (direita) da área de estudo.

Outros dados auxiliares foram coletados para auxiliar no processo de mapeamento. Porém, estes dados não foram considerados na construção do conjunto de treinamento para a área de estudo, a exemplo dos dados de precipitação e de focos de incêndio. Como os dados de precipitação do satélite *Tropical Rainfall Measuring Mission,* TRMM (KUMMEROW *et al.,* 2000) apresentam uma resolução espacial muito baixa, de cerca de 30 km, toda a área de estudo só correspondia a seis pixels, não influenciando na resposta da classificação dos polígonos caso fossem utilizados como atributo preditivo. Caso semelhante ocorreu com os dados referentes aos focos de incêndio, obtidos através do mapeamento de queimadas do INPE, uma vez que só foram encontrados apenas quinze polígonos com ocorrência de fogo para os dois anos.

5.3. Imagens *Landsat*

Foram utilizadas duas imagens *TM Landsat-5* (resolução espacial de 30 m) referentes à órbita/ponto 220/74, adquiridas em 01/06/2006 e 27/07/2009, como mostra a Figura 5.6. Como primeira etapa, as imagens foram georreferenciadas usando como referência a base de dados da NASA (*Global Land Cover Facility* - GLCF).

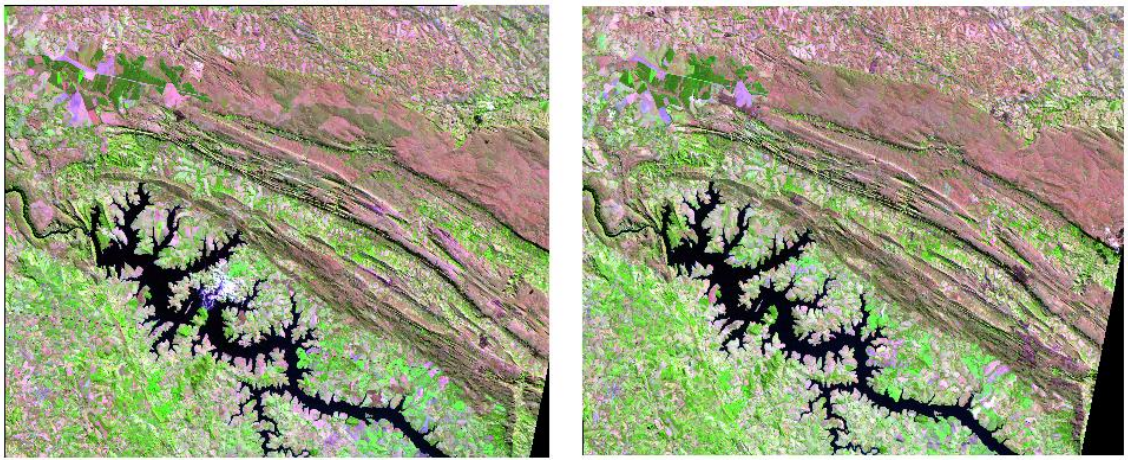


Figura 5.6 Imagens *Landsat* TM-5 (R5B4G3) de 2006 (esquerda) e 2009 (direita).

No processamento, foram consideradas as bandas azul (1), verde (2), vermelho (3), além de três bandas de infravermelho (4, 5 e 7) para extração de informações espectrais.

5.4. Imagens MODIS

Como o bioma apresenta duas estações bem definidas, esta característica pode auxiliar na detecção de mudanças nas coberturas vegetais. Para analisar esta situação, foram utilizadas as imagens do sensor MODIS, com correção atmosférica e resolução espacial de 250 m, para os períodos seco e úmido da região e os dados temporais dos índices de vegetação EVI e EVI2. Especificamente, foram utilizados perfis anuais como vetores de características para 2006 e 2009. Cada perfil anual foi extraído da composição EVI e EVI2 de 16 dias do MODIS, resultando em 23 instantes para cada perfil (365 dias dividido por 16). A Figura 5.7 ilustra um resumo da sequência de imagens obtidas dos dados MODIS para os 23 instantes do ano de 2009. Como descrito no capítulo 3, para cada imagem, cada pixel corresponde a um valor de índice de vegetação (EVI ou EVI2, para este estudo), de modo que o empilhamento desta sequência de imagens resulta em uma série temporal.

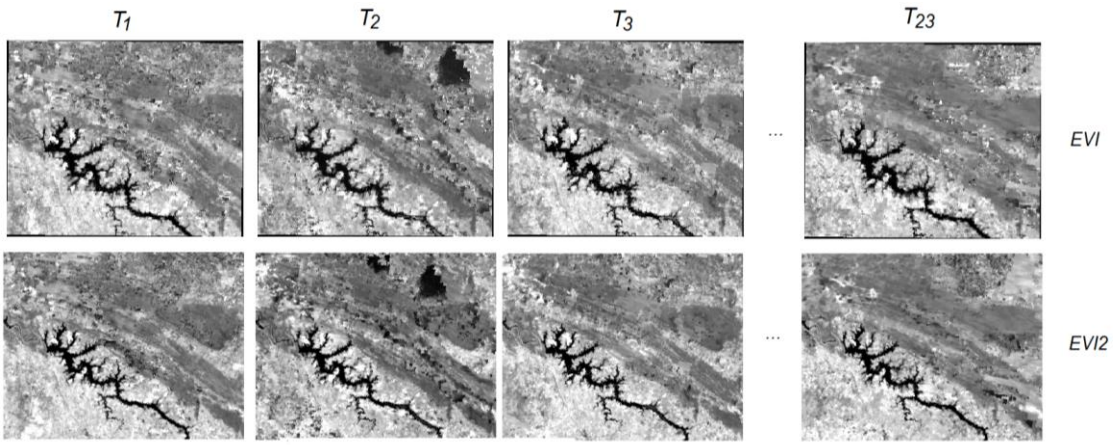


Figura 5.7 Sequência de imagens para compor os perfis anuais de EVI e EVI2 para o ano de 2009.

5.5. Transformação

As áreas de campo nativo apresentam árvores esparsas, cujas sombras podem ser identificadas nas imagens *Landsat*. Portanto, a técnica de Modelo Linear de Mistura Espectral (MLME) pode ser utilizada para decompor a imagem em componentes de solo, sombra e vegetação, e avaliar se a componente sombra pode ser utilizada como atributo para separar as classes. Além disso, como muitas pastagens cultivadas estão em processo de degradação, pode ocorrer a presença de solo exposto em algumas destas áreas. Com isso, também levou-se em consideração o uso do componente solo como atributo preditivo.

Desta forma, as componentes de solo, sombra e vegetação foram geradas a partir da transformação das imagens *Landsat* (R5B4G3) de 2006 e 2009. Para a geração das imagens-fração, utilizou-se o módulo de processamento de imagens (TerraPDI) implementadas na TerraLib (CÂMARA *et al.,* 2008), como *plugin* do *TerraView* (INPE, 2013). O *TerraView* é um aplicativo para visualizar dados geográficos com recursos de consulta a análise destes dados. Para gerar os componentes, fez-se a seleção dos *endmembers* e a escolha do método de combinação entre transformação de principais componentes e mínimos quadrados como estimador. A Figura 5.8 apresenta as componentes

do MLME geradas a partir da transformação da imagem *Landsat* do ano de 2006.

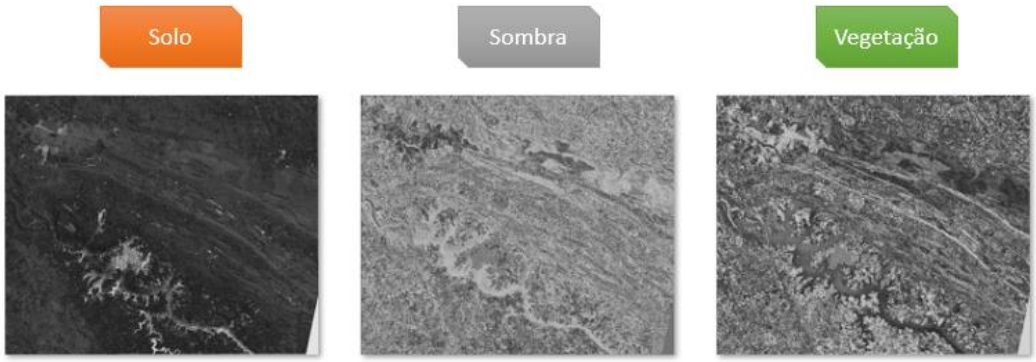


Figura 5.8 Componentes do MLME relativos ao ano de 2006.

5.6. Extração de atributos

Para a extração dos atributos preditivos, cada amostra foi considerada como o conjunto de informações relativas a um polígono do conjunto de regiões do *shapefile,* criado a partir dos mapas de referência. Os atributos foram pré-processados, conforme descrito no capítulo 4, e foram transformados por meio da normalização. Os atributos espaciais de cada polígono foram extraídos usando o sistema GeoDMA (KORTING *et al*., 2013). GeoDMA (*Geographic Data Mining Analyst*) é uma plataforma de código aberto que possui implementadas técnicas de segmentação de imagens, extração e seleção de atributos, métodos de análise multitemporal para detecção de mudanças e alguns tipos de classificação (árvores de decisão, redes neurais) e mineração de dados espaciais.

Os atributos espaciais extraídos para cada polígono foram: área; altura, largura e área do retângulo envolvente (envelope); valor do ângulo principal; área do menor círculo circunscrito; compacidade, densidade, dimensão fractal; índice da forma de cada polígono; razão entre a área do polígono e a mínima elipse circunscrita; perímetro, e razão entre o perímetro e a área do polígono. Os cálculos de cada um desses atributos estão descritos em Korting (2012).

O sistema GeoDMA foi utilizado, também, para extrair os atributos espaciais e espectrais relativos à altitude e à declividade a partir dos dados TOPODATA e os atributos espectrais a partir das imagens *Landsat* (bandas 1-5 e 7) para os anos 2006 e 2009. Os atributos espectrais extraídos são: amplitude, dissimilaridade, entropia, homogeneidade, média, moda, razão, desvio padrão e soma dos pixels de cada polígono (KORTING, 2012).

Em relação às imagens MODIS, para cada uma das 23 imagens dos índices EVI e do EVI2, os atributos relacionados com os índices de vegetação foram extraídos usando a biblioteca TerraLib. Foi necessário implementar um algoritmo para associar a média dos valores de todos os pixels (no caso, os índices de vegetação) para cada um dos polígonos. Os atributos dos índices de vegetação correspondem, assim, à média dos valores dos pixels dentro de cada polígono, para cada imagem EVI e EVI2 no instante T_i (i = 1,2, ...,23).

Para evitar o problema de pixels de borda, foi utilizada uma nova abordagem para definir os atributos de índices de vegetação para cada polígono. Usando a biblioteca TerraLib, implementamos um algoritmo que define o valor do atributo EVI (Imagens T_1 a T_{23}) para cada polígono, equivalente à média dos "pixels puros" do polígono, conforme ilustrado na Figura 5.9.

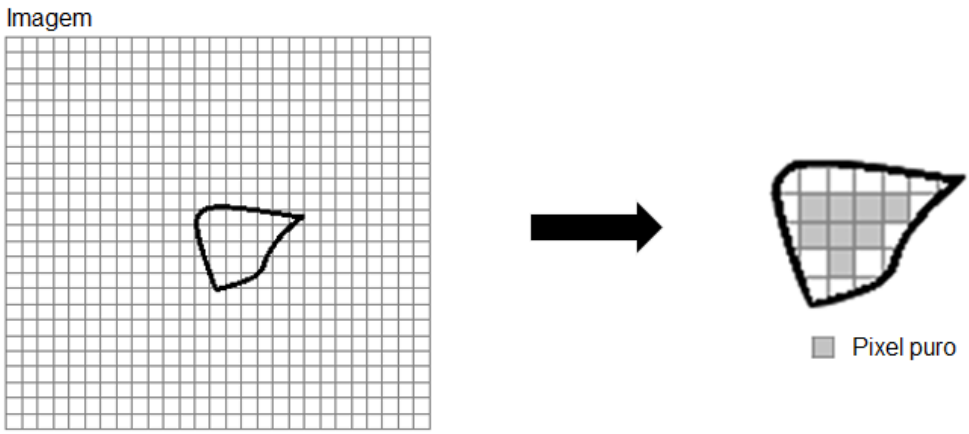


Figura 5.9 Abordagem de pixels puros.

Nesta abordagem, foi considerado que no cálculo da média apenas os “pixels puros”, ou seja, os pixels cujos seus vizinhos também pertencem à mesma classe, evitando, assim, que pixels de borda influenciem no cálculo da média dos valores de EVI e EVI2 de um dado polígono. A partir desta estratégia, consequentemente, polígonos muito pequenos, que estão inclusos em único pixel são desconsiderados do conjunto de amostras, conforme a Figura 5.10. A mesma abordagem foi usada no cálculo dos atributos extraídos das imagens-fração solo e sombra.

Figura 5.10 Exemplo de polígono desconsiderado do conjunto de amostras na abordagem de pixels puros.

Ao final do último experimento, cada amostra continha 134 atributos – o atributo classe e os seguintes atributos preditivos:

- 13 atributos espaciais obtidos a partir dos mapas de referência: área; altura, largura e área do retângulo envolvente; valor do ângulo principal; área do menor círculo circunscrito; compacidade; densidade; dimensão fractal; índice da forma de cada polígono; razão entre a área do polígono e a mínima elipse circunscrita; perímetro, e razão entre o perímetro e a área do polígono.

- Atributos espectrais dos dados TOPODATA (18) e das imagens *Landsat* (54): para cada banda (altitude e declividade do TOPODATA; e bandas 1-5 e 7 da imagem *Landsat*), foram utilizados os atributos amplitude, dissimilaridade, entropia, homogeneidade, média, moda, razão, desvio padrão e soma.

- 46 atributos relativos à média dos índices de vegetação MODIS: bandas EVI referentes aos instantes 1 a 23, assim como os atributos correspondentes às 23 bandas EVI2.

- 2 atributos extraídos do MLME: componente solo e componente sombra.

5.7. Treinamento e classificação

Foi utilizada a estratégia de divisão percentual do conjunto de total de amostras, sendo 2/3 das amostras para o conjunto de treinamento e 1/3 para o conjunto de teste, além de realizar o balanceamento da quantidade de amostras de cada classe. Os seguintes métodos de classificação foram usados nos experimentos: árvores de decisão, florestas aleatórias e SVM. Para a classificação baseada em árvores de decisão foram utilizados o GeoDMA e o ambiente Weka (HALL *et al.*, 2009), que utilizam o algoritmo C4.5 para criação do modelo de classificação e seleção de atributos. O Weka foi usado também para a construção do modelo de classificação baseado no método de florestas aleatórias. Para o método SVM, utilizou-se a biblioteca LibSVM (CHANG; LIN, 2011), que disponibiliza ferramentas para particionar um conjunto de dados em treinamento e teste, treinar o classificador e fazer a predição do conjunto de teste, além de estimar os parâmetros C e γ do modelo e *kernel* do classificador.

5.8. Avaliação

Na fase de avaliação dos resultados, foram realizadas comparações e análises das predições dos classificadores. No processo de avaliação, as matrizes de confusão e os índices Kappa (COHEN, 1960) foram usados. A partir da matriz

de confusão, pode-se calcular medidas estatísticas descritivas, como a acurácia dos resultados, para que a confusão entre classes possa ser claramente apresentada e entendida (CONGALTON; GREEN, 2008).

O *software* Weka foi usado para gerar as matrizes de confusão e acurácia global dos algoritmos de árvores de decisão e florestas aleatórias, e o índice Kappa global para análise da predição das amostras do conjunto de teste. No caso da classificação SVM, os cálculos da matriz de confusão e do índice Kappa global foram implementados em linguagem R.

6. RESULTADOS E DISCUSSÕES

Para a geração e análise dos resultados, 4 experimentos foram realizados. Cada experimento usa um conjunto de atributos no processo de classificação, como mostra a Tabela 6.1, onde cada conjunto de dados é representado pelo conjunto de atributos distintos descritos na seção 5.6 para cada tipo de dado utilizado na metodologia.

Tabela 6.1 – Resumo dos conjuntos de dados utilizados em cada experimento.

Experimentos	**Dados utilizados**
1	EVI
	Landsat
	EVI + Landsat
	TOPODATA
	EVI + Landsat + TOPODATA
2	EVI (pixels puros)
	Landsat (pixels puros)
	EVI + Landsat (pixels puros)
	TOPODATA (pixels puros)
	EVI + Landsat + TOPODATA (pixels puros)
3	MLME (pixels puros)
	EVI + TOPODATA + Landsat + MLME (pixels puros)
4	EVI + SOLO (MLME) (pixels puros)

As classes de interesse utilizadas neste estudo são: Pastagem Cultivada, Campo Rupestre, Campo (Campo Limpo/Campo Sujo) e Outros. Três métodos de classificação foram usados nos experimentos: árvores de decisão, SVM e florestas aleatórias. Os parâmetros do método SVM foram ajustados para cada teste, usando a técnica de busca exaustiva com validação cruzada. Para o método de florestas aleatórias, a quantidade de árvores geradas pelo classificador foi fixada em 10. O índice Kappa foi calculado para todos os resultados de classificação com o intuito de mensurar a concordância entre os resultados dos classificadores e o conjunto de amostras de referência.

No processo de classificação do Experimento 1, primeiramente avaliou-se o potencial das imagens EVI como único conjunto de atributos na classificação, a fim de analisar a contribuição das séries temporais de índices de vegetação na classificação. Ademais, vale ressaltar que o uso das imagens EVI2 no processo de classificação também foi avaliado, mas os testes apresentaram resultados similares aos das imagens EVI. Portanto, os experimentos apresentados aqui utilizaram o atributo EVI. O conjunto de dados contém um total de 8.000 amostras (o conjunto de informações relativas a um polígono), e foi utilizada a estratégia de divisão percentual de 2/3 das amostras para o conjunto de treinamento e 1/3 para o conjunto de teste. A Tabela 6.2 mostra os resultados do Experimento 1.

Tabela 6.2 – Resultados do Experimento 1 para 4 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI	55,11	0,4014	Árvores de decisão
	64,50	0,5267	SVM
	63,79	0,5172	Florestas aleatórias
Landsat	47,35	0,2972	Árvores de decisão
	49,94	0,3325	SVM
	47,54	0,2999	Florestas aleatórias
EVI + Landsat	58,93	0,4524	Árvores de decisão
	64,42	0,5256	SVM
	59,89	0,4650	Florestas aleatórias

Tabela 6.2 – Conclusão.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
TOPODATA	51,36	0,3504	Árvores de decisão
	49,50	0,3267	SVM
	50,62	0,3412	Florestas aleatórias
EVI + Landsat + TOPODATA	57,28	0,4287	Árvores de decisão
	66,83	0,5577	SVM
	61,95	0,4926	Florestas aleatórias

Em seguida, foi avaliada a contribuição dos atributos relativos às informações espaciais e espectrais das imagens Landsat, descritos nas seções 5.3 e 5.6. Os resultados utilizando apenas estas informações mostraram que a maioria dos atributos selecionados através da ferramenta Weka para a construção dos modelos eram relativos às bandas infravermelho (4 e 5). No entanto, somente o uso destes atributos não implicou em uma boa classificação, como pode ser visto na Tabela 6.2. Posteriormente, foi analisado se a combinação destes dois tipos de atributos (EVI e Landsat) alcançaria resultados melhores, porém foram obtidas acurácias e índices Kappa semelhantes aos obtidos por meio dos dados EVI.

Analogamente aos testes com EVI e com Landsat, analisou-se o conjunto de atributos relativos aos dados TOPODATA se utilizado sozinho para classificação. Contudo, os resultados coletados mostraram que eles não apresentaram uma boa classificação. Em seguida, foi realizada a combinação dos três tipos de atributos (EVI, Landsat e TOPODATA) e, a partir da avaliação dos resultados, pode-se perceber um pequeno aumento na acurácia para os algoritmos de árvores de decisão e SVM em relação aos outros testes

aplicados. Como ilustrado na Tabela 6.2, o melhor resultado da classificação utilizou o método baseado em SVM e atingiu acurácia de apenas 66,83% e índice Kappa de 0,5577.

Como pode ser visto na Tabela 6.3, a matriz de confusão para os dados EVI, resultante do modelo de classificação para árvores aleatórias, indica que muitas amostras de Campo foram classificadas como Campo Rupestre, e vice-versa. Devido à confusão na discriminação entre estas duas classes, duas classificações para cada experimento foram realizadas. A primeira classificou as regiões em 4 classes: Pastagem Cultivada, Campo Rupestre, Campo e Outros. A segunda classificou as amostras em 3 classes: Pastagem Cultivada, Campo Nativo e Outros. As classes Campo e Campo Rupestre foram agrupadas em uma única classe, nomeada Campo Nativo. Desta forma, o classificador identificou 3 classes com um total de 12.000 amostras, utilizando também a técnica de divisão percentual para particionar os conjuntos de treinamento e teste.

Tabela 6.3 – Matriz de confusão para o Experimento 1 (EVI com florestas aleatórias).

Class./Ref.	**Campo**	**Campo rupestre**	**Pastagem cultivada**	**Outros**
Campo	**374**	*125*	34	104
Campo Rupestre	*160*	**371**	14	140
Pastagem cultivada	26	18	**619**	33
Outros	132	138	61	**371**

Então, uma segunda classificação foi empregada nas amostras do Experimento 1, como pode ser visto na Tabela 6.4. Para 3 classes, os resultados para todos os testes apresentaram uma melhora bastante satisfatória para as acurácias e os índices ao comparar com a classificação para 4 classes.

Tabela 6.4 – Resultados do Experimento 1 para 3 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI	67,47	0,5112	Árvores de decisão
	75,32	0,6298	SVM
	74,89	0,6232	Florestas aleatórias
Landsat	61,11	0,4174	Árvores de decisão
	66,48	0,4972	SVM
	63,73	0,4565	Florestas aleatórias
EVI + Landsat	73,04	0,5956	Árvores de decisão
	78,65	0,6798	SVM
	74,79	0,6221	Florestas aleatórias
TOPODATA	63,50	0,4538	Árvores de decisão
	63,51	0,4526	SVM
	62,16	0,4330	Florestas aleatórias
EVI + Landsat + TOPODATA	72,70	0,5905	Árvores de decisão
	79,76	0,6965	SVM
	77,03	0,6555	Florestas aleatórias

O método SVM caracterizou novamente a melhor classificação, obtendo uma acurácia de 79,76% e Kappa de 0,6965. Para analisar a contribuição dos atributos neste experimento, o modelo de árvore de decisão com atributos EVI, Landsat e TOPODATA, é ilustrado na Figura 6.1. Pode-se observar que, ao analisar a influência dos atributos TOPODATA, partindo do nó raiz, as regiões

classificadas como Pastagem Cultivada encontram-se em um ramo do modelo, com a média da altitude abaixo de 900,53 m, enquanto que as amostras com valores acima desse valor foram classificadas como Campo, Campo Rupestre e Outros. Esta informação condiz com a altitude das fitofisionomias de Campo Rupestre descritas em Ribeiro e Walter (2008).

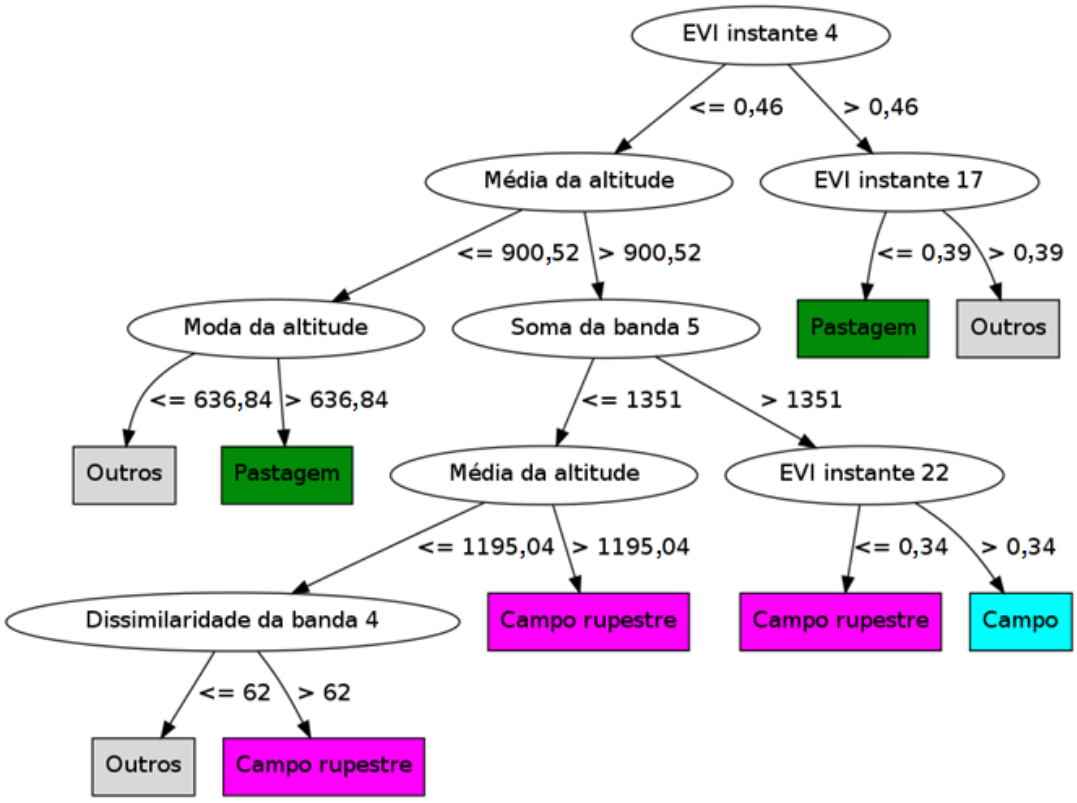


Figura 6.1 Árvore de decisão para o Experimento 1 (EVI + Landsat + TOPODATA).

Com o mesmo conjunto de atributos do primeiro experimento, realizou-se o Experimento 2, aplicando um novo enfoque na definição dos valores dos atributos preditivos relacionados aos índices de vegetação. Para cada amostra, o valor de cada atributo representa apenas a média dos “pixels puros” (conforme seção 5.6) do polígono representado pela amostra. As Tabelas 6.5 e 6.6 exibem os resultados da classificação para 4 e 3 classes, respectivamente.

Tabela 6.5 – Resultados do Experimento 2 para 4 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI (pixels puros)	73,99	0,5871	Árvores de decisão
	74,38	0,5713	SVM
	77,46	0,6387	Florestas aleatórias
Landsat (pixels puros)	56,07	0,3062	Árvores de decisão
	59,61	0,2868	SVM
	62,43	0,3708	Florestas aleatórias
EVI + Landsat (pixels puros)	71,68	0,5217	Árvores de decisão
	68,97	0,5146	SVM
	72,83	0,5588	Florestas aleatórias
TOPODATA (pixels puros)	60,12	0,3237	Árvores de decisão
	61,58	0,3885	SVM
	60,12	0,3399	Florestas aleatórias
EVI + Landsat + TOPODATA (pixels puros)	76,88	0,6281	Arvores de decisão
	70,44	0,5465	SVM
	76,88	0,6257	Florestas aleatórias

Tabela 6.6 – Resultados do Experimento 2 para 3 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI (pixels puros)	85,55	0,7579	Árvores de decisão
	81,28	0,6902	SVM
	87,28	0,7816	Florestas aleatórias
Landsat (pixels puros)	63,58	0,4047	Árvores de decisão
	59,46	0,3898	SVM
	68,21	0,4525	Florestas aleatórias
EVI + Landsat (pixels puros)	78,03	0,6370	Árvores de decisão
	79,56	0,6672	SVM
	82,08	0,6961	Florestas aleatórias
TOPODATA (pixels puros)	69,36	0,4556	Árvores de decisão
	68,97	0,4514	SVM
	70.52	0.4730	Florestas aleatórias
EVI + Landsat + TOPODATA (pixels puros)	84,97	0,7478	Arvores de decisão
	78,57	0,6430	SVM
	84,97	0,7465	Florestas aleatóras

Pode-se observar claramente que, a partir dos resultados obtidos com a abordagem de pixels puros, as classificações implicaram em resultados melhores do que utilizando todos os pixels (Experimento 1). A melhora no resultados ocorreu devido a uma provável adição de ruído dos pixels de borda dos polígonos, os quais são considerados no cálculo dos valores dos atributos

do conjunto de dados do Experimento 1. Contudo, como vários polígonos foram desconsiderados através desta nova abordagem e, consequentemente, a quantidade de amostras foi reduzida. O conjunto de atributos relativos ao dados EVI passou a ter um total de 508 amostras com pixels puros, por exemplo. Tanto a classificação para 4 como para 3 classes sofreram um aumento superior a 10% na acurácia para alguns testes. Esses testes para classificações em 3 classes, por exemplo, obtiveram taxas de acerto acima de 80% e índices Kappa maiores que os obtidos no Experimento 1.

Como ilustrado na Tabela 6.5, os melhores resultados da classificação para 4 classes utilizaram os métodos baseados em árvores de decisão e florestas aleatórias, com todas as fontes de dados disponíveis (EVI, Landsat e TOPODATA), atingindo a acurácia de 76,88% e índices Kappa de 0,6281 e 0,6257. E conforme a Tabela 6.6, o melhor resultado da classificação para 3 classes utilizou o método de florestas aleatórias com os dados EVI, atingindo a acurácia de 87,28% e índice Kappa de 0,7816. Pode-se observar que utilizando todas as fontes de dados foram obtidos resultados semelhantes, atingindo a acurácia de 84,97% para o algoritmo de árvores de decisão.

No Experimento 3 foram incluídos os atributos relativos às imagens-fração de solo e sombra do MLME. Vale ressaltar que a abordagem de pixels puros também foi utilizada neste experimento, uma vez que, por meio da avaliação dos Experimentos 1 e 2, a abordagem com pixels puros resultou em acurácias melhores do que observando todos os pixels de cada amostra. Os resultados são apresentados nas Tabelas 6.7 e 6.8.

Tabela 6.7 – Resultados do Experimento 3 para 4 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
Solo + Sombra (MLME) (pixels puros)	51,84	0,3067	Árvores de decisão
EVI + Landsat + TOPODATA + MLME (pixels puros)	75,14	0,6148	
Solo + Sombra (MLME) (pixels puros)	50,94	0,2953	SVM
EVI + Landsat + TOPODATA + MLME (pixels puros)	71,43	0,5586	
Solo + Sombra (MLME) (pixels puros)	47,23	0,2395	Florestas aleatórias
EVI + Landsat + TOPODATA + MLME (pixels puros)	77,46	0,6414	

Tabela 6.8 – Resultados do Experimento 3 para 3 classes.

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
Solo + Sombra (MLME) (pixels puros)	58,16	0,3736	Árvores de decisão
EVI + Landsat + TOPODATA + MLME (pixels puros)	87,28	0,7863	
Solo + Sombra (MLME) (pixels puros)	57,22	0,3627	SVM
EVI + Landsat + TOPODATA + MLME (pixels puros)	80,30	0,6738	
Solo + Sombra (MLME) (pixels puros)	53,62	0,2997	Florestas aleatórias
EVI + Landsat + TOPODATA + MLME (pixels puros)	86,13	0,7631	

Pode-se observar que, similarmente ao Experimento 1, apenas os dados de solo e sombra não contribuíram para melhorar os resultados da classificação. Ademais, a integração destes atributos gerou respostas com acurácias similares aos resultados anteriores para 4 classes. Já para 3 classes, o experimento obteve acurácia de 87,28% por meio da classificação baseada em

árvores de decisão, utilizando todas as fontes de dados (EVI, Landsat, TOPODATA e MLME).

Ao introduzir o componente solo no conjunto de atributos, este teve peso importante na construção da árvore de decisão e no mapeamento deste resultado, como pode ser ilustrado na Figura 6.2, com o nó do atributo do componente solo ocupando a posição raiz no modelo da árvore. Este modelo gerou um dos melhores resultados da classificação, com acurácia de 87,28% e índice Kappa de 0,7863. Pode-se observar que, a maioria dos atributos selecionados na construção da árvore são relativos ao EVI, mostrando a importância da informação temporal na classificação.

A partir da análise dos melhores atributos da árvore de decisão supracitada, foi realizado o Experimento 4, considerando apenas a integração de dados EVI e do componente Solo do MLME com 508 amostras, já que estes dados foram preponderantes na criação da maioria dos modelos apresentados nos experimentos. A Tabela 6.9 exibe os resultados da classificação para 4 e 3 classes, respectivamente.

Tabela 6.9 – Resultados do Experimento 4 para 3 e 4 classes.

Dados	***N. de classes***	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI + SOLO (pixels puros)	4	77,46	0,6499	Árvores de decisão
		77,59	0,6530	SVM
		79,77	0,6720	Florestas aleatórias
	3	87,28	0,7897	Árvores de decisão
		83,25	0,7208	SVM
		90,17	0,8324	Florestas aleatórias

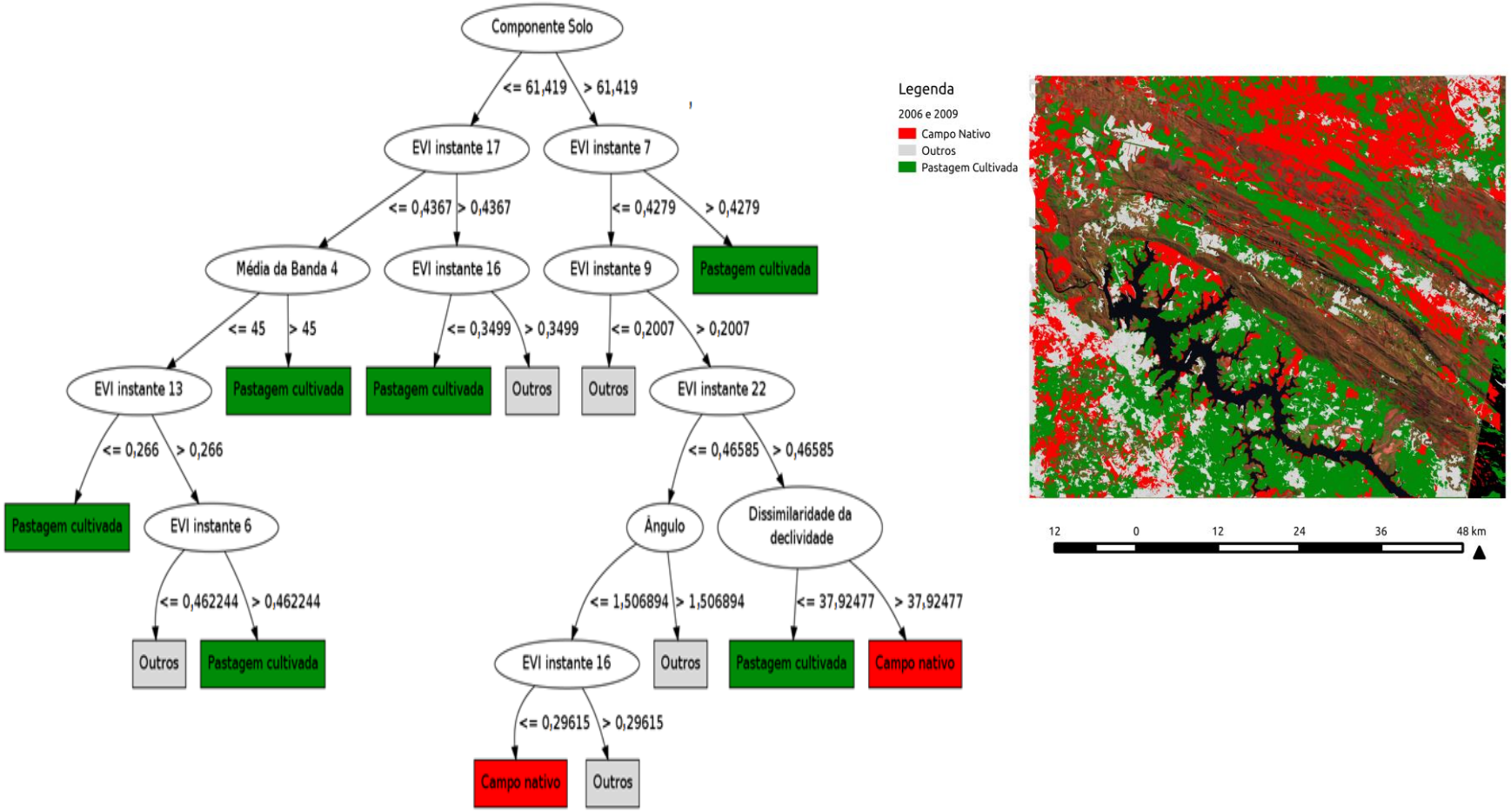


Figura 6.2. Árvore de decisão (à esquerda) e mapeamento (à direita) para o melhor resultado encontrado no Experimento 3.

Com isso, pode-se observar que o último experimento gerou modelos com as melhores acurácias para os três classificadores. Especificamente, o classificador de florestas aleatórias gerou o melhor resultado da classificação para 3 classes, com acurácia de 90,17% e índice Kappa de 0,8374. A matriz de confusão resultante da classificação para este experimento, apresentando um conjunto total de 508 amostras particionadas para treinamento e teste, está exposta na Tabela 6.10.

Tabela 6.10 – Matriz de confusão do Experimento 4 (3 classes – Florestas aleatórias).

Class./Ref.	**Campo Nativo**	**Pastagem cultivada**	**Outros**
Campo	**51**	6	0
Pastagem cultivada	0	**87**	3
Outros	5	3	**18**

Com os 4 experimentos, realizou-se a comparação dos principais resultados alcançados, observando a contribuição de todos os conjuntos de dados utilizados, para cada um dos classificadores, conforme exibido na Tabela 6.11.

Tabela 6.11 – Comparação dos resultados mais relevantes (3 Classes).

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI (pixels puros)	85,55	0,7579	Árvores de decisão
EVI (pixels puros)	87,28	0,7816	Florestas aleatórias
EVI + Landsat + TOPODATA + MLME (pixels puros)	86,13	0,7631	Florestas aleatórias
EVI + Landsat + TOPODATA + MLME (pixels puros)	87,28	0,7863	Árvores de decisão
EVI + SOLO (pixels puros)	87,28	0,7897	Árvores de decisão
EVI + SOLO (pixels puros)	90,17	0,8374	Florestas aleatórias

Em relação ao custo computacional, o algoritmo SVM apresentou um tempo de processamento ligeiramente maior do que os classificadores. Quando utilizado sem o ajuste de parâmetros, o SVM apresenta o mesmo tempo de processamento comparado às árvores de decisão e florestas aleatórias, realizando as predições em questão de segundos. Porém, a fase de estimação dos parâmetros C e γ mais adequados foi a principal responsável por este custo adicional, acrescentando um tempo médio de 15 minutos para classificação para 4 classes e cerca de 1 hora para 3 classes, para cada um dos testes.

A Tabela 6.11 mostra que os resultados da classificação usando todos os atributos são similares ao da classificação com apenas os atributos EVI. Do ponto de vista operacional, a criação de um modelo de classificação por árvores de decisão com os dados EVI e o componente solo do MLME seria a opção com menor custo computacional para o mapeamento das 3 classes. A opção pelo algoritmo de florestas aleatórias demandaria maior custo computacional, visto que este algoritmo gera um número maior de árvores no modelo.

Todos os algoritmos, de forma geral, tiveram resultados satisfatórios no caso da classificação das 3 classes, para discriminar Campo Nativo de Pastagem Cultivada. Conforme a Tabela 6.8, a integração de todos os atributos resultou em valores de acurácia superiores a 80% e índices Kappa no intervalo de 0,7 a 0,84 (com índice máximo de 0,7863). Além disso, a análise e seleção dos melhores atributos a partir da observação das árvores de decisão do Experimento 3, relativos aos dados EVI e à componente solo do MLME, resultou em um valor de acurácia de 90,17% e Kappa de 0,8324.

Por outro lado, a discriminação entre as formações campestres nativas, separando o Campo Rupestre da classe Campo (Campo Limpo/Campo Sujo), não obteve bons resultados, conforme mostra a Tabela 6.12. O máximo valor de acurácia obtido neste caso foi de 79,77%.

Tabela 6.12 – Comparação dos resultados mais relevantes (4 Classes).

Dados	***Acurácia (%)***	***Kappa***	***Algoritmo***
EVI + Landsat + TOPODATA	79,76	0,6965	SVM
EVI (pixels puros)	77,46	0,6387	Florestas aleatórias
EVI + Landsat + TOPODATA + MLME (pixels puros)	77,46	0,6414	Florestas aleatórias
EVI + SOLO (pixels puros)	77,59	0,6530	SVM
EVI + SOLO (pixels puros)	79,77	0,6720	Florestas aleatórias

7. CONCLUSÕES E TRABALHOS FUTUROS

O objetivo deste trabalho foi classificar áreas de formações campestres nativas (Campo Limpo, Campo Sujo e Campo Rupestre) e regiões de pastagens cultivadas no Cerrado brasileiro. A estratégia de análise usada para resolver este problema foi integrar dados de diferentes fontes e usar técnicas de processamento de imagens e mineração de dados no processo de classiifcação. Para isso, uma região do Parque da Serra da Canastra, situada em Minas Gerais, foi utilizada como área de estudo, uma vez que podem ser encontradas áreas correspondentes às classes de interesse nestas localizações.

O processo de análise de imagens e mineração de dados foi realizado com auxílio dos ambientes computacionais GeoDMA e Weka, e da biblioteca LibSVM. Alguns algoritmos foram implementados na biblioteca TerraLib e em linguagem R para auxiliar no processo de classificação e validação dos resultados. No tocante ao tempo de processamento, o algoritmo SVM apresentou um custo mais alto para classificação dos resultados em relação aos outros dois classificadores, sendo que o fator predominante para este custo foi a estimação dos parâmetros mais adequados.

Foram realizadas dois tipos de classificação: uma mais genérica, para separar as classes de campo nativo das de pastagem cultivada; e outra que procurava além da separação de classes antrópicas das nativas, uma divisão mais detalhada das formações campestres nativas.

Três experimentos foram conduzidos, combinando atributos preditivos de diferentes fontes. Informações espaciais das amostras (polígonos), dados de topografia, imagens de média resolução espacial (*Landsat* TM-5), índices de vegetação provenientes do sensor MODIS e imagens-fração de solo e sombra derivadas do modelo de mistura espectral foram utilizadas para a extração dos atributos preditivos, o treinamento e a avaliação do conjunto de amostras.

Os algoritmos de árvores de decisão, SVM e florestas aleatórias, de forma geral, apresentaram bons resultados na separação entre campo nativo e pastagens cultivadas. Com isso, conclui-se que a integração de dados de diferences tipos e resoluções, juntamente com técnicas de mineração de dados, constituem numa metodologia adequada para o mapeamento de classes Campo Nativo e Pastagem Cultivada no Cerrado Brasileiro. Os melhores resultados obtidos tiveram acurácias superiores a 85% e índices Kappa na faixa de 0,6 a 0,84. Pode-se ainda destacar o componente solo obtido pelo modelo de mistura, quando integrado apenas com o conjunto de atributos preditivos relativos ao EVI, resultou no melhor resultado da classificação, com acurácia de 90,17% e índice Kappa de 0,8374.

Ao tentar realizar uma classificação mais detalhada, tentando discriminar as áreas de campo nativo em seus subtipos e a partir destas fazer a separação com pastagens cultivadas, verificou-se uma redução na acurácia dos resultados. Com isto, nota-se que uma classificação ótima entre as quatro classes, na prática, é bastante difícil de ser obtida devido à similaridade espectral entre as classes.

Como trabalhos futuros, imagens de alta resolução podem ser investigadas, uma vez que elas foram utilizadas apenas na identificação das classes e interpretação visual. A inclusão de dados de campo também pode ser útil para o mapeamento das classes assim como dados auxiliares que possam contribuir no processo de discriminação entre as classes.

Além disso, do ponto de vista de aprimoramento de ferramentas de análise imagens, há interesse na construção de novos classificadores de imagens no *plugin* GeoDMA. Pretende-se, posteriormente, implementar os métodos de SVM e de florestas aleatórias no GeoDMA para melhorar a identificação de padrões de mudança no uso e cobertura do solo do Cerrado brasileiro e nas demais aplicações.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHÃO, A.; MIGLIORANZA, E.; GODOY JR., M. Geração de imagens de proporções através de um modelo linear de mistura. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 6. (SBSR), 1990, Manaus. **Anais...** São José dos Campos: INPE, 1990. p. 146-151. Printed, On-line. ISBN 978-85-17-00051-5. (INPE-5170-PRE/1639). Disponível em: <http://urlib.net/dpi.inpe.br/marte@80/2008/08.15.12.46>. Acesso em: 03 abr. 2014..

ADAMS, J. B.; KAPOS, V.; SMITH, M. O.; ALMEIDA FILHO, R.; GILLESPIE, A. R.; ROBERTS, D. A. A new Landsat view of land use in Amazonia. In: INTERNATIONAL SYMPOSIUM ON PRIMARY DATA ACQUISITION, 1990, Manaus, BR. **Proceedings...** 1990. p. 177-179. (INPE-7850-PRE/3690).

AGUIAR, A. P. D. **Utilização de atributos derivados de proporções de classes dentro de um elemento de resolução de imagem ("pixel") na classificação multiespectral de imagens de sensoriamento remoto**. 1991. 227 p. (INPE-5306-TDI/456). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), Sao Jose dos Campos, 1991. Disponível em: <http://urlib.net/6qtX3pFwXQZ3r59YD6/GNVHb>. Acesso em: 03 abr. 2014.

BEDDINGTON, J. Food, energy, water and the climate: A perfect history of global events? **Lecture to Sustainable development UK 09**, p. 1-9, 2009.

BIAS, E. D. S.; BARBOSA, F. L. R.; BRITES, R. S. Emprego de imageamento passivo na análise da variabilidade espacial da turbidez no espelho d'água do Lago Paranoá, Distrito Federal. **Eng. Sanit. Ambient.**, v. 18, n. 1, p. 55-64, 2013.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 5., 1992, Pittsburgh, Pennsylvania, USA. **Proceedings…** Pennsylvania: ACM, 1992. p. 144-152.

BOVOLO, F.; CAMPS-VALLS, G.; BRUZZONE, L. A support vector domain method for change detection in multitemporal images. **Pattern Recognition Letters**, v. 31, n. 10, p. 1148-1154, 2010.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BROSSARD, M.; BARCELLOS, A. O. Conversão do cerrado em pastagens cultivadas e funcionamento de latossolos. **Cadernos de Ciência & Tecnologia**, Brasília, v. 22, n. 1, p. 153-168, 2005.

BRUZZONE, L.; SMITS, P.; TILTON, J. Foreword special issue on analysis of multitemporal remote sensing images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 41, n. 11, p. 2419-2422, nov. 2003.

CÂMARA, G.; VINHAS, L.; FERREIRA, K.; QUEIROZ, G.; SOUZA, R.; MONTEIRO, A.; CARVALHO, M.; CASANOVA, M.; FREITAS, U. TerraLib: An open source GIS library for large-scale environmental and socio-economic applications. **Open Source Approaches in Spatial Data Handling.**, Berlin, v. 2, p. 247-270, 2008.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. **ACM Transaction on Intelligent Systems and Technology**, v. 2, n. 3, p. 1-27, 2011.

CHAVES, J. M.; MOREIRA, L.; SANO, E. E.; BEZERRA, H. S.; FEITOZA, L. Uso da técnica de segmentação na identificação dos principais tipos de pastagens cultivadas do Cerrado. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 10. (SBSR), 2001, Foz do Iguaçu. **Anais...** São José dos Campos: INPE, 2001. p. 31-33. CD-ROM, On-line. ISBN 85-17-00016-1. Disponível em:<http://urlib.net/dpi.inpe.br/lise/2001/09.12.16.36>. Acesso em: 03 abr. 2014.

COHEN, J. A. Coefficiente of agreement for nominal scales. **Educational and Psychological Measurement**, n. 20, p. 37-46, 1960.

COLLINS, J. B.; WOODCOCK, C. E. An Assessment of Several Linear Change Detection Techniques for Mapping Forest Mortality Using Multitemporal Landsat TM Data. **Remote Sensing of Environment**, v. 56, n. 1, p. 66-77, 1996.

COLSTOUN, E. C. B. D.; STORY, M. H.; THOMPSON, C.; COMMISSO, K.; SMITH, T. G.; IRONS, J. R. National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. **Remote Sensing of Environment**, v. 85, n. 3, p. 316-327, 2003.

CONGALTON, R. G.; GREEN, R. **Assessing the accuracy of remotely sensed data:** principles and practices. 2. ed. Boca Raton: CRC Press, p. 1-183, 2008.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Maching Learning**, v. 20, n. 3, p. 273-297, set. 1995.

COSTA, G. A. O. P. D. **A knowledge-based approach for automatic interpretation of multidate remote sensing data**. 2009. 149 p. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2009

COSTA, M. H.; PIRES, G. F. Effects of Amazon and Central Brazil deforestation scenarios on the duration of the dry season in the arc of deforestation. **International Journal of Climatology**, v. 30, n. 13, p. 1970-1979. 2010.

COURA, S. M. D. C. **Mapeamento de vegetação do estado de Minas Gerais utilizando dados MODIS**. 2006. 150 p. (INPE-14657-TDI/1213). Dissertação de Mestrado - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2006. Disponível em:<http://urlib.net/sid.inpe.br/MTC-m13@80/2006/12.21.13.36>. Acesso em: 03 abr. 2014.

DING, C. H. Q.; DUBCHAK, I. Multi-class protein fold recognition using support vector machines and neural networks. **Bioinformatics**, v. 17, n. 4, p. 349-358, 2001.

EITEN, G. **Delimitação do conceito de cerrado**. Rio de Janeiro: Arquivos do Jardim Botânico, v. 21, p. 125-134, 1977.

EKLUNDH, L.; JÖNSSON, P. **TIMESAT 3.1 software manual**. Lund, Sweden: Malmö University, 2012. 182 p.

EMPRESA BRASILEIRA DE PESQUISAS AGROPECUÁRIAS (EMBRAPA)/ INSTITUTO NACIOANLA DE PESQUISAS ESPACIAIS (INPE). **Levantamento de informações de uso e cobertura da terra na Amazônia** - sumário executivo. Brasília/São José dos Campos:TerraClass. 2011.

FAYYAD, U.; SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, New York, NY, v. 39, n. 11, p. 27-34, 1996.

FERREIRA, L. G.; FERNANDEZ, L. E.; SANO, E. E.; FIELD, C.; SOUSA, S. B.; ARANTES, A. E.; ARAÚJO, F. M. Biophysical Properties of Cultivated Pastures in the Brazilian Savanna Biome: An Analysis in the Spatial-Temporal Domains Based on Ground and Satellite Data. **Remote Sensing**, v. 5, n. 1, p. 307-326, 2013.

FERREIRA, L. G.; SANO, E. E.; FERNANDEZ, L. E.; ARAÚJO, F. M. Biophysical characteristics and fire occurrence of cultivated pastures in the Brazilian savanna observed by moderate resolution satellite data. **International Journal of Remote Sensing**, v. 34, n. 1, p. 154-167, 2013.

FERREIRA, L. G.; YOSHIOKA, H.; HUETE, A.; SANO, E. E. Seasonal landscape and spectral vegetation index dynamics in the Brazilian Cerrado: An analysis within the Large-Scale Biosphere–Atmosphere Experiment in Amazônia (LBA). **Remote Sensing of Environment**, v. 87, n. 4, p. 534-550, 2003.

FREITAS, R. M.; SHIMABUKURO, Y. E. Combining wavelets and linear spectral mixture model for MODIS satellite sensor time-series analysis. **Journal of Computational Interdisciplinary Sciences**, v. 1, p. 51-56, 2008.

FREITAS, S. R.; CRUZ, C. B. M. Análise de componentes principais e modelo linear de mistura na discriminação de classes de vegetação na Mata Atlântica. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12. (SBSR), 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. p. 1529-1536. CD-ROM, On-line. ISBN 85-17-00018-8. Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2004/11.04.11.55>. Acesso em: 03 abr. 2014.

FRIELD, M. A.; BRODLEY, C. E.; STRAHLER, A. H. Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales. **IEEE Transactions on Geoscience and Remote Sensing**, v. 37, n. 2, p. 969-977, 1999.

GHIMIRE, B.; ROGAN, J.; GALIANO, V. R.; PANDAY, P.; NEETI, N. An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. **GIScience & Remote Sensing**, v. 49, n. 5, p. 623-643, 2012.

GHIMIRE, B.; ROGAN, J.; MILLER, J. Contextual Land-Cover Classification: Incorporating Spatial Dependence in Land-Cover Classification Models Using Random Forests and the Getis Statistic. **Remote Sensing Letters**, v. 1, n. 1, p. 45-54, 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 1, n. 1, 2009.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. 2. ed. San Francisco, CA: Morgan Kaufmann Publishers, 2006, 743 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN. **The elements of statistical learning:** data mining, inference, and predicition. Stanford, CA: Springer, 2009, 745 p..

HEAS, P.; DATCU, M. Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning. **IEEE Transaction on Geoscience and Remote Sensing**, v. 43, n. 7, p. 1635-1647, jul. 2005.

HUETE, A. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, v. 25, n. 3, p. 295-309, 1988.

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

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Mapas de biomas do Brasil**, 2004. Disponivel em: <http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=169>. Acesso em: 27 fev. 2013.

_____. **Manual técnico da vegetação brasileira**. 2. ed. Rio de Janeiro: Manuais Técnicos em Geociências, v. 1, 2012.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **TerraView**. São José dos Campos, SP: Divisão de Processamento de Imagens, 2013. Disponivel em: <www.dpi.inpe.br/terraview>. Acesso em: 26 Jan. 2014.

JIANG, Z.; HUETE, A. R.; DIDAN, K.; MIURA, T. Development of a two-band enhanced vegetation index without a blue band. **Remote Sensing of Environment**, v. 112, n. 10, p. 3833-3845, 2008.

JUSTICE, C. O.; TOWNSHEND, J. R. G.; VERMOTE, E. F.; MASUOKA, E.; WOLFE, R. E.; SALEOUS, N.; ROY, D. P.; MORISETTE, J. T. An overview of MODIS Land data processing and product status. **Remote Sensing of Environment**, v. 83, n. 1, p. 3-15, 2002.

KLINK, C.; MACHADO, R. Conservation of the Brazilian Cerrado. **Conservation Biology**, v. 19, n. 3, p. 707-713, jun. 2005.

KORTING, T. S. **GeoDMA:** a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery. 2012. 119p. (sid.inpe.br/mtc-m19/2012/07.31.18.22-TDI). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012. Disponível em: <http://urlib.net/8JMKD3MGP7W/3CCH86S>. Acesso em: 03 abr. 2014.

KORTING, T. S.; FONSECA, L. M. G.; CÂMARA, G. GeoDMA - Geographic Data Mining Analyst: a framework for GIScience. **Computers & Geosciences**, v. 57, p. 133-145, 2013. doi: <10.1016/j.cageo.2013.02.007>.

KUMMEROW, C.; SIMPSON, J.; THIELE, O.; BARNES, W.; CHANG, A. T. C.; STOCKER, E.; ADLER, R. F.; HOU, A.; KAKAR, R.; WENTZ, F. et al. The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. **Jorunal of Applied Methodology**, v. 39, p. 1965-1982, 2000.

LAMBIN, E.; LINDERMAN, M. Time series of remote sensing data for land change science. **IEEE Transactions on Geoscience and Remote Sensing**, v. 44, n. 7, p. 1926-1928, 2006.

LEITE, P. B. C.; FEITOSA, R. Q.; FORMAGGIO, A. R.; COSTA, G. A. O. P. D.; PAKZAD, K.; SANCHES, I. D. Hidden Markov Models for crop recognition in remote sensing image sequences. **Pattern Recognition Letters**, v. 32, n. 1, p. 19-26, 2011.

LI, X.; YEH, A. G. O. Principal component analysis of stacked multitemporal images for the monitoring of rapid urban expansion in the Pearl River Delta. **International Journal of Remote Sensing**, v. 19, n. 8, p. 1501-1518, 1998.

LIU, X.; LATHROP, R. G. Urban change detection based on an artificial neural network. **International Journal of Remote Sensing**, v. 23, n. 12, p. 2513-2518, 2002.

LU, D.; MAUSEL, P.; BRONDÍZIO, E.; MORAN, E. Change detection techniques. **International Journal of Remote Sensing**, v. 25, n. 12, p. 2365-2401, 2004.

MACHADO, R. B.; RAMOS NETO, M.; PEREIRA, P.; CALDAS, E.; GONÇALVES, N.; SANTOS, K.; TABOR, K.; STEININGER, M. **Estimativa de perda da área do Cerrado Brasileiro**. Relatório Técnico. Conservação Internacional, Brasília, DF. 2004.

MALHADO, A. C. M.; PIRES, G. F.; COSTA, M. H. Cerrado Conservation is Essential to Protect the Amazon Rainforest. **Ambio**, v. 39, n. 8, p. 580-584, 2010.

MAS, J.; PALACIO, J.; PUIG, H.; SOSA-LOPEZ, A. Modelling deforestation using GIS and artificial neural networks. **Environmental Modelling and Software**, v. 19, p. 461-471, 2004.

MASCARENHAS, N. D. D.; CORREIA, V. R. M. Medidas de qualidade de estimadores de proporções de classes dentro de um pixel de imagens de satélite. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE PROGRESSO PARA AS CIÊNCIAS, 1982,Campinas. **Resumos…** São Paulo: SBPC, 1982. (INPE-2791-PRE/354).

MCCAULEY, S.; GOETZ, S. Mapping residential density pattern using multi-tempora Landsat data and a decision-tree classifier. **International Journal of Remote Sensing**, v. 25, n. 6, p. 1077-1094, mar. 2004.

MELGANI, F.; BRUZZONE, L. Classification of hyperspectral remote sensing images with support vector machines. **IEEE Transaction on Geoscience and Remote Sensing,** v. 42, n. 8, p. 1778-1790, 2004.

MELLO, M. P. D.; SILVA, G. B. S. D.; PEREIRA, G.; PRADO, B. R. D.; SHIMABUKURO, Y. E.; RUDORFF, B. F. T.; VIEIRA, C. A. O.; PETERNELLI, L. A. Avaliação do uso do modelo linear de mistura espectral na discriminação de fitofisionomias do cerrado. In: SIMPÓSIO NACIONAL DO CERRADO, 9.; SIMPÓSIO INTERNACIONAL DE SAVANAS TROPICAIS, 2., 2008, Brasília. **Anais...** Planaltina-DF: EMBRAPA CERRADOS, 2008. CD-ROM.

MELNYCHUK, A. **Multi-temporal crop classification using a decision tree in a southern ontario agricultural region**. 2012, 78 p. Dissertação (Mestrado em Geografia). University of Guelph, Guelph, Ontario, Canada, 2012.

MILLER, H. J.; HAN, J. **Geographic data mining and knowledge discovery**. Boca Raton, FL: CRC Press, 2009, 445 p.

NOMA, A.; KORTING, T. S.; FONSECA, L. M. G.; PAPA, J. P. Uma comparação entre classificadores usando regiões e perfis EVI para agricultura. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 16. (SBSR), 2013, Foz do Iguaçu. **Anais...** São José dos Campos: INPE, 2013. p. 2250-2257. DVD, Internet. ISBN 978-85-17-00066-9 (Internet), 978-85-17-00065-2 (DVD). Disponível em:<http://urlib.net/3ERPFQRTRW34M/3E7GDT8>. Acesso em: 03 abr. 2014.

NOVACK, T.; RIBEIRO, B. M. G.; KUX, H. J. H. Análise dos dados do satélite WorldView-2 para a discriminação de alvos urbanos semelhantes com base em algoritmos de seleção de atributos. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 15. (SBSR), 2011, Curitiba. **Anais...** São José dos Campos: INPE, 2011. p. 7815-7821. DVD, Internet. ISBN 978-85-17-00056-0 (Internet), 978-85-17-00057-7 (DVD). Disponível em:<http://urlib.net/3ERPFQRTRW/3A2L5KH>. Acesso em: 03 abr. 2014.

PAL, M. Random Forest Classifier for Remote Sensing Classification. **International Journal of Remote Sensing**, v. 26, n. 1, p. 217-222, 2005.

PEREIRA, J. L. G. **Estudos de áreas de floresta em regeneração através de imagens Landsat**. 1996. 157 p. (INPE-5987-TDI/578). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), Sao Jose dos Campos, 1996. Disponível em:<http://urlib.net/6qtX3pFwXQZ3r59YD6/GPcwz>. Acesso em: 03 abr. 2014.

PINHO, C. M. D.; FONSECA, L. M. G.; KORTING, T. S.; ALMEIDA, C. M.; KUX, H. J. H. Land-cover classification of an intra-urban environment using high-resolution images and object-based image analysis. **International Journal of Remote Sensing**, v. 33, n. 19, p. 5973-5995, 2012.

PINHO, C.; SILVA, F.; FONSECA, L.; MONTEIRO, A. Intra-urban land cover classification from high-resolution images using the C4.5 algorithm. In: ISPRS CONGRESS, 21., 2008, Beijin. **Proceedings…** Beijin: ISPRS, 2008. v.7.

QUILAN, J. R. **C4.5:** programs for machine learning. San Mateo, CA: Morgan Kaufmann, v. 1, 1993.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81-106, 1986.

RATTER, J. A.; RIBEIRO, J. F.; BRIDGEWATER, S. The Brazilian Cerrado and threats to its biodiversity. **Annals of Botany**, 80, 1997. 223-230.

RIBEIRO, J. F.; WALTER, B. M. T. As principais fitofisionomias do Bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P.; RIBEIRO, J. F. **Cerrado:** ecologia e flora. Brasília: EMBRAPA, v. 1, p. 152-212, 2008.

ROBERTS, D. A.; GARDNER, M.; CHURCH, R.; USTIN, S.; SCHEER, G.; GREEN, R. Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models. **Remote Sensing of Environment**, v. 65, n. 3, p. 267-279, 1998.

RODRIGUES, R. R.; GANDOLFI, S. Recomposição de florestas nativas: princípios gerais e subsídios para uma definição metodológica. **Revista Brasileira de Horticultura Ornamental**, Campinas, v. 2, n. 1, p. 4-15, 2001.

RODRIGUEZ-GALIANO, V. F.; GHIMIRE, B.; ROGAN, J.; CHICA-OLMO, R.-S. J. P. An Assessment of the effectiveness of a random forest classifier for land-cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 67, n. 1, p. 93-104, 2012.

RUSHING, J.; RAMACHANDRAN, R.; NAIR, U.; GRAVES, S.; WELCH, R.; LIN, H. ADaM: a data mining toolkit for scientists and engineers. **Computers and Geosciences**, v. 31, n. 5, p. 607-618, jun. 2005.

SANO, E. E.; BARCELLOS, A. O.; BEZERRA, H. S. Assessing the spatial distribution of cultivated pastures in the Brazilian savanna. **Pasturas Tropicales**, v. 22, n. 3, p. 2-15, 2001.

SANO, E. E.; ROSA, R.; LUÍS, J.; BRITO, S.; FERREIRA, G. Mapeamento semidetalhado do uso da terra do Bioma Cerrado. **Pesquisa Agropecuária Brasileira**, Brasília, v. 43, n. 1, p. 153-156, 2008.

SCARIOT, A.; SILVA, J. C. S.; FELFILI, J. M. **Cerrado:** ecologia, biodiversidade e conservação. Brasília: Ministério do Meio Ambiente, 2005, 439 p.

SCOLFORO, J. R.; MELLO, J. M.; OLIVEIRA, A. D. **Inventário florestal de Minas Gerais:** cerrado - florística, estrutura, diversidade, similaridade, distribuição diamétrica e de altura, volumetria, tendências de crescimento e áreas aptas para manejo florestal. Lavras: UFLA, 2008.

SESNIE, S. E.; FINEGAN, B.; GESSLER, P. E.; THESSLER, S.; BENDANA, Z. R.; SMITH, A. M. S. The Multispectral Separability of Costa Rican Rainforest Types with Support Vector Machines and Random Forest Decision Trees. **International Journal of Remote Sensing**, v. 31, n. 11, p. 2885-2909, 2010.

SHIMABUKURO, Y. E. **Shade images derived from linear mixing models of multispectral measurements of forested areas**. Tese (Doctor of Philosophy). Colorado State University, Fort Collins, CO. 1987.

SHIMABUKURO, Y. E.; MELLO, E. M. K.; MOREIRA, J. C.; DUARTE, V. **Segmentação e classificação da imagem sombra do modelo de mistura para mapear desflorestamento na Amazônia**. São José dos Campos: INPE, 1997.

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

SILVA, G. B. S. D.; FORMAGGIO, A. R.; SHIMABUKURO, Y. E.; ADAMI, M.; SANO, E. E. Discriminação da cobertura vegetal do Cerrado matogrossense por meio de imagens MODIS. **Pesq. agropec. bras.**, Brasília, v. 45, n. 2, p. 186-194, 2010.

SILVA, G. B. S. D.; SPINELLI, L. D. A.; NOGUEIRA, S. F.; BOLFE, E. L.; VICTORIA, D. D. C.; VICENTE, L. E.; GREGO, C. R.; ANDRADE, R. G. Sistema de Informação Geográfica (SIG) e base de dados geoespaciais do projeto GeoDegrade. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 16. (SBSR), 2013, Foz do Iguaçu. **Anais...** São José dos Campos: INPE, 2013. p. 2487-2493. DVD, Internet. ISBN 978-85-17-00066-9 (Internet), 978-85-17-00065-2 (DVD). Disponível em: <http://urlib.net/3ERPFQRTRW34M/3E7GJK4>. Acesso em: 03 abr. 2014.

SKOLE, D. L.; , C. W. H.; SALAS, W. A.; NOBRE, C. A. Physical and human dimensions of deforestation in Amazonia. **Biosciences**, v. 44, n. 5, p. 314-322, 2012.

SMITH, A. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. **Journal of Spatial Science**, v. 55, n. 1, p. 69-79, 2010.

SUNG, A. H.; MUKKAMALA, S. **Identifying important features for intrusion detection using support vector machines and neural networks**. In: SYMPOSIUM ON APPLICATIONS AND THE INTERNET, 2003,Washington. **Proceedings…** Washington, DC, USA: IEEE Computer Society. 2003. p. 209-216.

TAN, P.-N.; STEINBACH, M.; KUMAR, V.; POTTER, C.; KLOOSTER, S.; TORREGROSA, A. Finding spatio-temporal patterns in earth science data. **Earth Science**, p. 1-12, 2001.

TEIXEIRA, C. G. **Validação do modelo linear de mistura espectral em imagens aster/terra a partir de dados ikonos**. 2004, 127p. (INPE-13183-TDI/1029). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004. Disponível em: <http://urlib.net/sid.inpe.br/jeferson/2005/02.15.15.35>. Acesso em: 03 abr. 2014.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. USA: Academic Press, 2008, 984 p.

TUCKER, C. J. Red and photographic infrared linear combinations for monitoring vegetation. **Remote Sensing of Environment**, v. 8, n. 2, p. 127-150, 1979.

TUCKER, C. J.; PINZON, J. E.; BROWN, M. E.; SLAYBACK, D. A.; PAK, E. W.; MAHONEY, R.; VERMOTE, E. F.; SALEOUS, N. E. An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. **International Journal of Remote Sensing**, v. 26, n. 20, p. 4485−4498, 2005.

VALERIANO, M. D. M. Modelo digital de variáveis morfométricas com dados SRTM para o território nacional: o projeto TOPODATA. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12. (SBSR), 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. p. 3595-3602. CD-ROM, On-line. ISBN 85-17-00018-8. (INPE-12739-PRE/8029). Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2004/10.29.11.41>. Acesso em: 03 abr. 2014.

VAPNIK, V. N. **The nature of statistical learning theory**. New York, NY, USA.: Springer-Verlag New York, Inc., 1995, 314 p.

WALTER, B. M. T. **Fitofisionomias do bioma Cerrado : síntese terminológica e relações florísticas.** 2006. 389p. Tese (Doutorado em Ecologia) - Universidade de Brasília, Brasília. 2006.

WEBB, A. R. **Statistical Pattern Recognition**. 2. ed. Chichester: John Wiley & Sons, 2002, 514 p.

WITTEN, I. H.; FRANK, E. **Data mining:** practical machine learning tools and techniques with Java implementations. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, 371 p.

WOODCOCK, C. E.; MACOMBER, S. A.; PAX-LENNEY, M.; COHEN, W. B. Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. **Remote Sensing of Environment**, v. 78, n. 1, p. 192-203, 2001.

ZHANG, X.; FRIEDL, M. A.; SCHAAF, C. B.; STRAHLER, A. H.; HODGES, J. C.; GAO, F.; REED, B. C.; HUETE, A. Monitoring vegetation phenology using MODIS. **Remote Sensing of Environment**, v. 84, n. 3, p. 471-475, 2003.