Ministério da
Ciência, Tecnologia e Inovação

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras.

Monografia referente a disciplina SER - 300 - Introdução ao Geoprocessamento, sob orientação do Prof. Dr. Antonio Miguel Vieira Monteiro.

Cesare Di Girolamo Neto - 130338

INPE

São José dos Campos.

Junho, 2014

ABSTRACT

Coffee is the main crop produced in the southern region of the state of Minas Gerais, Brazil, and techniques for estimating the area used for this crop are being intensely investigated in order to produce reliable yield estimates. Coffee trees have a similar spectral pattern to forest, making it difficult to automatically classify these land use types. The application of Random Forest algorithm in order to perform the classification of remote sense data is a promising approach in discriminating more complex classes of land use/cover. This study presents an application of Random Forest for the automatic classification of remote sensing data, which was adapted for the identification of land use with emphasis in the coffee areas in the Machado, South region of Minas Gerais state in Brazil. The software used for preprocessing the data and validate the ratings was the SPRING and the WEKA. For this region, the methodology for the application of Random Forest was divided in three main stages: in the first, the pre-processing of the data was accomplished. Grey level masks were created in each one of the 11 bands of the images from the LANDSAT-8 satellite; in the second stage, the Random Forest was trained and applied on the image in order to verify its potential for discriminating coffee areas; the third stage consisted of the analysis and validation of the results, using as reference the map classified visually. The Kappa index and the accuracy of the models generated were used to select the best classified map. The result of the best classification by Random Forest was 84.13% with a Kappa index of 0.6.The Random Forest algorithm presented good results when compared to previous studies. A few bands from the LANDSAT-8 proved to be useless on the classification.

Keywords : Random Forest, data mining, automatic classification, coffee maping.

Sumário:

1 Introdução

A análise da cobertura da terra consiste em determinar como uma área de interesse é utilizada, permitindo a caracterização das interações antrópicas com o meio ambiente. O conhecimento da dinâmica de transformação do uso da terra mostra-se cada vez mais importante, afim de analisar a forma pela qual determinado espaço está sendo ocupado (PRUDENTE e ROSA, 2007). Essa análise pode ser usada como suporte às decisões de planejamento e ao desenvolvimento sustentável, uma vez que o espaço esta atrelado às necessidades e atividades humanas (SANTOS e PETRONZIO, 2011).

No Brasil, grandes transformações em seu espaço foram decorrentes do desenvolvimento do setor agrícola (HELFAND e RESENDE, 2000). A preocupação com o uso da terra e a cobertura vegetal vem crescendo rapidamente, pois, as formas como as quais esses dois fatores estão sendo manipulados influenciam o modo de vida da população (CANDIDO et al., 2010). Com levantamentos do uso e cobertura da terra, os padrões de organização do espaço podem ser compreendidos e pode-se observar diversas conseqüências, como o uso inadequado do solo (ROSA, 2009).

Para obtenção e análise dos padrões do uso da terra, mostra-se importante atrelar dados advindos de sensores remotos (BRANNSTROM et al., 2008). Esta tecnologia permite gerar mapas de uso do solo, os quais têm grande importância por demonstrarem, a partir da interpretação de imagens de satélites, áreas ocupadas por pastagem, culturas agrícolas, vegetação natural, cursos de rios e outras feições.

Neste sentido, a cultura do café, embora esta apresente variações no comportamento espectral (captado por satélites), por causa de fatores como espaçamento, idade, época do ano, ela pode ser identificada e mapeada em imagens de satélites de média resolução espacial, com boa precisão de mapeamento, desde que o analista realize uma interpretação visual sobre os resultados da classificação feita no computador (MOREIRA et al., 2004). O café pode ser mapeado por meio de imagens de satélite e os resultados podem ser disponibilizados tanto espacialmente como tabelados por macrorregião, microrregião, município e Estado (MOREIRA et al., 2007).

O mapeamento de áreas cafeeiras torna-se ainda mais atrativo dado que o Brasil é atualmente o maior produtor de café do mundo, sendo que em 2012 foi responsável por cerca

de 37% da produção mundial, o equivalente a 55,9 milhões de sacas de 60 Kg. Com cerca de 60% da produção destinada ao mercado externo, os ganhos anuais do país com a exportação deste grão chegaram próximos a US$ 6 bilhões (USDA, 2013). O café é cultivado em 12 estados brasileiros, sendo que o maior produtor é Minas Gerais, com cerca de 51% da produção nacional (MINISTÉRIO DA AGRICULTURA, 2013).

O estado de Minas Gerais teve sua área cafeeira mapeada por Moreira et al. (2007). Este trabalho visou mapear todo o estado e identificar áreas cafeeiras e não cafeeiras para imagens do satélite Landsat 5 do ano de 2006. Os resultados mostraram que a região sul de minas é responsável por cerca de 50% da área de café plantada do estado, ressaltando ainda mais sua importância no contexto nacional. Dentre outras conclusões obtidas pelos autores está a dificuldade de identificação de áreas cafeeiras com áreas de cerrado e eucalipto. De maneira geral, a cultura cafeeira apresenta resposta espectral bastante complexa, em função da variabilidade de seus diversos parâmetros, tais como declividade, espaçamento entre plantas, estado vegetativo, estágio fenológico, sombreamento e manejo, entre outros (VIEIRA et al., 2006).

A metodologia convencional de levantamento e atualização de informações sobre o uso da terra é, normalmente, caracterizada pelo alto custo e dificuldade na obtenção de dados, o que pode limitar sua aplicação. Avanços computacionais que auxiliam a extração de informação e conhecimento de imagens de sensoriamento remoto, bem como o uso e sistemas de informação geográfica (SIG) se mostram essenciais para armazenar, analisar e apresentar os mais variados tipos de informação georreferenciada (BURROUGH e MCDONNELL, 1998).

A metodologia computacional de mineração de dados demonstra ter alto potencial de aplicação em estudos relacionados ao mapeamento de áreas cafeeiras, sendo utilizados diversas técnicas de classificação, como redes neurais artificiais (ANDRADE et al., 2011; 2013) e maquinas de vetores suporte (BISPO et al., 2014). realizaram um procedimento análogo para a região de Machado (MG). Não existe a melhor técnica, cada uma possui vantagens e desvantagens. A escolha de uma técnica requer uma análise detalhada do problema em questão e a decisão de qual representação e estratégia de descoberta é a mais adequada.

O crescente uso por uma técnica de mineração de dados chamada de florestas aleatórias, ou, *Random Forest*, tem chamado a atenção em estudos relacionados a identificação do uso da terra (RODRIGUEZ-GALIANO et al., 2012). Esta técnica também pode ser

considerada uma das mais precisas quando comparada a outras, como redes neurais artificiais e máquinas de vetores suporte (CARUANA et al., 2008). As florestas aleatórias ainda são computacionalmente muito efetivas, além de evitarem sobreajuste (*overfitting*) e serem pouco sensíveis a ruídos (BREIMAN, 2001). Estas observações evidenciam a importância de pesquisas capazes de gerar informações que possam ser utilizadas para fornecer suporte ao mapeamento de áreas cafeeiras por métodos computacionais.

Sendo assim, este trabalho visa investigar métodos computacionais para análise da geoinformação e disponibilização de conhecimento para estudos relacionados à cafeicultura. A hipótese deste trabalho é que o uso da técnica de Random Forest é um método eficiente, com melhor desempenho do que outras técnicas de mineração de dados, para o mapeamento de áreas plantadas com a cultura do café.

2 Objetivos

2.1 Objetivo geral

Este trabalho teve como objetivo avaliar o processo de classificação automática utilizando *Random Forest* para identificação de áreas cafeeiras em imagens do satélite LANDSAT-8 para o município de Machado/MG.

2.2 Objetivos específicos

Os objetivos específicos foram:

- Verificar se *Random Forest* é uma técnica adequada para a classificação automática de áreas cafeeiras a partir de imagens de satélite;

- Avaliar o desempenho desta técnica a partir da variação de seus parâmetros internos e custo computacional.

- Comparar o desempenho de *Random Forest* com Árvores de decisão.

- Avaliar as melhores bandas espectrais para classificação automática de áreas cafeeiras.

3 Revisão bibliográfica

3.1 Sensoriamento Remoto e dados de satélite

Na literatura são encontradas várias definições do que é sensoriamento remoto. Uma delas é que sensoriamento remoto é uma ciência que obtém informações de um determinado objeto, área ou fenômeno sem o contado direto com os objetos investigados (LILLESAND, 1987). Outra definição é que sensoriamento remoto é a utilização de modernos sensores, aeronaves, espaçonaves com o objetivo de estudar o ambiente terrestre por meio do registro e análise das interações da radiação eletromagnética (REM) com alvos na superfície terrestre (NOVO, 1989).

A radiação eletromagnética (REM) que interage com estes alvos, pode ser absorvida, refletida, transmitida e emitida seletivamente (MOREIRA, 2001). Com o desenvolvimento tecnológico atual, é possível medir, com razoável precisão, as propriedades espectrais desses alvos. O uso de sistemas sensores em nível orbital, para obtenção de dados da radiação refletida e/ou emitida pelos alvos da superfície terrestre, é muito importante para o reconhecimento da superfície da Terra de maneira rápida e eficaz. No caso da vegetação, a REM pode ser absorvida, refletida ou transmitida, podendo cada uma ser hemisférica ou bidirecional (PONZONI e SHIMABUKURO, 1991).

A REM absorvida pelos pigmentos da folha é, em parte, dissipada em forma de calor ou fluorescência, sendo que apenas uma pequena parcela desta energia é armazenada devido a fotossíntese. Já a parte da REM refletida pelas folhas pode ser influenciada por três principais fatores, sendo eles: quantidade e características de pigmentos, quantidade de espaços preenchidos pela água, e morfologia celular (JENSEN, 2011).

Para observar estes fatores podem ser utilizadas uma imagem proveniente de um sensor remoto, ou também chamada de cena. Cenas coletadas pelos sensores podem ser obtidas por diferentes categorias de satélite, como os militares, científicos e meteorológicos (MOREIRA 2003). Os sensores utilizados para dados em nível orbital estão geralmente voltados para os estudos dos recursos naturais e operam em diferentes números de bandas que estão associadas aos comprimentos de onda. Neste trabalho foram utilizadas informações obtidas dos sensores presentes no satélite LANDSAT-8.

3.1.1 O satélite LANDSAT-8

O programa Landsat foi desenvolvido pela NASA (National Aeronautics and Space Administration) no inicio dos anos 70, com o objetivo de coletar dados sobre recursos naturais renováveis e não-renováveis da superfície terrestre (NOVO, 1989). Neste programa foram lançados ao menos 8 satélites, sendo que o mais recente é o LANDSAT-8.

Lançado do dia 11 de fevereiro de 2013, o LANDSAT-8, apresenta uma órbita circular com 98,2º de inclinação a uma altitude de 705 km (EMBRAPA, 2014). Existem dois sensores a bordo do satélite LANDSAT-8, o *Operational Land Imager* (OLI) e o *Thermal Infrared Sensor* (TIRS).

O sensor OLI opera em 9 faixas espectrais, também chamadas bandas espectrais do espectro eletromagnético, que correspondem a comprimentos de onda específicos de cada sensor. A resolução espacial destas de 30 metros, com exceção para a banda pancromática, a qual tem resolução espacial de 15m. Já o sensor TIRS opera em 2 faixas espectrais com resolução espacial de 100m. Estas e demais informações sobre estes sensores podem ser encontradas na Tabela 1.

As imagens obtidas por sensores remotos em diferentes canais são produzidas em escalas de cinza, sendo que a quantidade de REM refletida pelos objetos vai determinar a sua representação nesses diferentes tons, entre o branco (quando refletem toda a energia) e o preto (quando absorvem toda a energia). As principais aplicações das bandas espectrais do satélite Landsat-8, de acordo com INPE (2014) e EMBRAPA (2014), são:

Banda 1 (0,433-0,453 µm) Costal: Projetada especificamente para investigação de recursos hídricos e zonas costeiras.

Banda 2 (0,450 - 0,515 µm) Azul: Útil para mapeamento de águas costeiras, diferenciação entre solo e vegetação, mapeamento de florestas e detecção de feições culturais (mancha urbana, rodovias, etc.), entre outras.

Banda 3 (0,525 - 0,600 µm) Verde: Apresenta grande sensibilidade à presença de pigmentos, possibilitando sua análise em termos de quantidade e qualidade. Corresponde à reflectância da vegetação verde e sadia.

Sensor	Bandas Espectrais	Resolução Espectral	Resolução Espacial	Resolução Temporal	Área Imageada	Resolução Radiométrica
OLI (Operational Land Imager)	Costal (B1)	0,433-0,453 µm	30m	16 dias	185Km	12 bits
	Azul (B2)	0,450-0,515 µm				
	Verde (B3)	0,525-0,600 µm				
	Vermelho (B4)	0,630-0,680 µm				
	Infravermelho próximo (B5)	0,845-0,885 µm				
	Infravermelho médio (B6)	1,560-1,660 µm				
	Infravermelho médio (B7)	2,100-2,300 µm				
	Pancromático (B8)	0,500-0,680 µm	15m			
	Cirrus (B9)	1,360-1,390 µm	30m			
TIRS (Thermal Infrared Sensor)	Infravermelho termal (B10)	10,30-11,30 µm	100m			
	Infravermelho termal (B11)	11,50-12,50 µm				

Tabela 1: Informações sobre os sensores OLI e TIRS do satélite LANDSAT-8.
Fonte: EMBRAPA (2014)

Banda 4 (0,630 - 0,680 µm) Vermelho: Útil para discriminação entre espécies de plantas e delinear solo e feições culturais. Permite um bom contraste entre áreas ocupadas com vegetação e aquelas sem vegetação, apresentando níveis de cinza mais escuros para áreas com vegetação e níveis mais claros para áreas descobertas. É a banda mais utilizada para delimitar manchas urbanas e identificar áreas agrícolas.

Banda 5 (0,845 - 0,885 µm) Infravermelho próximo: É útil para identificação de culturas agrícolas, enfatizando a diferenciação solo/agricultura e água/solo. Apresenta sensibilidade à morfologia do terreno, permitindo o mapeamento de corpos d'água como rios, lagos e reservatórios.

Bandas 6 e 7 (1,560-1,660 µm e 2,100-2,300 µm) Infravermelho médio: Apresenta sensibilidade ao teor de umidade das plantas, permitindo detectar estresse na vegetação causado pela falta de água. Sensível à morfologia do terreno. Importante para estudos nas áreas de Geomorfologia, Solos e Geologia.

Banda 9 (1,360-1,390 µm) cirrus: Utilizada para a detecção de cirrus (nuvens formadas na alta atmosfera).

Bandas 10 e 11 (10,30-11,30 µm e 11,50-12,50 µm) Infravermelho termal: Apresenta sensibilidade nos fenômenos relativos aos contrastes térmicos. Usada para estudos de propriedades termais de rochas, solos, vegetação e água. Também utilizada para mapeamento da temperatura de águas oceânicas superficiais.

3.2 Processo de descoberta de conhecimento

Com a popularização da internet nos anos 90 e os avanços tecnológicos nas áreas de coleta, armazenamento e transmissão de grandes volumes de dados, o mundo se encontrou em situação "rica em dados, mas pobre em informação" (HAN et al., 2011). À medida que se buscava trabalhar com esses dados, percebeu-se que havia uma desproporção entre a quantidade de dados gerados e capacidade de analisar os dados, tornando-se inviável uma análise manual. Uma forma automatizada de tratar os dados seria um avanço valioso e traria vantagens competitivas consideráveis (FRAWLEY et al., 1992).

A demanda por geração de uma técnica computacional e ferramentas para análise de dados originou o campo da descoberta de conhecimento em bases de dados (KDD - *Knowledge Discovery in Databases*). Inicialmente, o KDD foi definido como a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados (FRAWLEY et al., 1992). Posteriormente, Fayyad et al. (1996) revisaram o conceito de KDD, sendo este redefinido como sendo o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados.

O processo KDD é composto de várias fases, sendo que para Fayyad et al. (1996), a mineração de dados (*data mining*) representa uma parte deste processo. Ela é a responsável pela extração de padrões embutidos em grandes volumes de dados, por meio da aplicação de algoritmos específicos. Uma aplicação imprudente de métodos de mineração de dados pode

ser uma atividade perigosa e pode conduzir a descoberta de padrões incorretos ou sem sentido (AGRAWAL et al., 1996). A Figura 1 representa uma visão geral do processo KDD, o qual é constituído das seguintes fases (FAYYAD et al., 1996):

- Fase de seleção: Dados são selecionados de acordo com critérios pré-definidos.
- Fase de pré-processamento: Neste estágio ocorre a limpeza dos dados, há a exclusão de informações desnecessárias, tratamento de dados ausentes e outras atividades.
- Fase de transformação: Os dados são configurados de forma a atender as exigências de uma dada técnica de mineração de dados, ocorre a conversão de dados e a derivação de novos atributos.
- Fase de mineração de dados: Extraem-se padrões de comportamento dos dados, a partir de uma técnica pré-determinada.
- Fase de interpretação: Os padrões são interpretados gerando conhecimento, os quais darão suporte à tomada de decisões humanas.

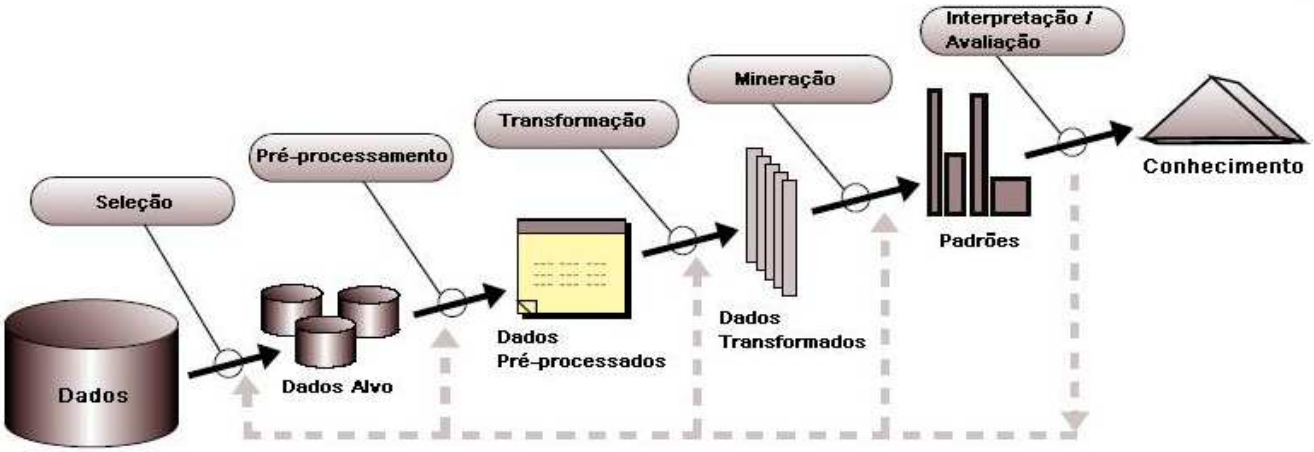

Figura 1: Fases do processo KDD

Fonte: Adaptado de FAYYAD et al. (1996).

Esse processo pode envolver várias iterações e quase sempre é necessário o retorno para fases anteriores.

3.2.1 Técnicas de mineração de dados

Na prática, os dois objetivos principais da mineração de dados são a predição e a descrição. A predição envolve o uso de variáveis com valores conhecidos para predizer um valor desconhecido ou futuro de outra variável (atributo meta). A descrição caracteriza propriedades gerais encontradas nos dados, com foco em padrões interpretáveis pelo ser humano. Esses objetivos podem ser alcançados por meio de vários tipos de tarefas e a escolha de uma ou mais destas depende do problema em questão (HAN et al., 2011; FAYYAD et al., 1996).

Cada tarefa de mineração de dados possui técnicas diferentes associadas, podendo haver abordagens híbridas, as quais aplicam duas ou mais técnicas em conjunto. Não existe a melhor técnica, cada uma possui vantagens e desvantagens. A escolha de uma técnica requer uma análise mais detalhada do problema em questão e a decisão de qual representação e estratégia de descoberta é a mais adequada.

As duas principais tarefas de mineração de dados são as descritivas e as preditivas. As tarefas descritivas procuram a identificação de padrões inerentes a determinado conjunto de dados, sendo que este conjunto não possui um atributo classe especificado. Já as tarefas preditivas têm como objetivo a construção de modelos, a partir de um determinado conjunto de dados, para posterior predição do comportamento de novos dados. As principais tarefas de predição são classificação e regressão.

A Classificação consiste na predição de um valor categórico ou discreto (atributo meta), e busca a construção de modelos e definição de regras, a partir de um conjunto de exemplos pré-classificados corretamente, para posterior classificação de exemplos novos e desconhecidos (REZENDE et al., 2002). A medida que a regressão visa descobrir uma função que mapeie um item de dados para uma variável de predição de valor numérico contínuo.

Neste trabalho a tarefa utilizada é a classificação, com ênfase em duas técnicas de aprendizado: árvores de decisão e florestas aleatórias.

3.2.1.1 Árvores de decisão

A indução de árvores de decisão é uma técnica de mineração de dados utilizada para descobrir regras de classificação para um atributo a partir da subdivisão dos dados em um conjunto que está sendo analisado. Árvores de decisão são simples representações de conhecimento e classificam exemplos em um número finito de classes (APTE e WEISS, 1997). Elas podem ser representadas graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido (WITTEN et al., 2011). Sua representação visual torna muito mais fácil para o usuário analisar e compreender os resultados (FAYYAD et al., 1996).

O nó raiz é o primeiro nó da árvore, no topo da estrutura. Os nós internos, incluindo o nó raiz, são nós de decisão. Cada um destes contém um teste sobre um ou mais atributos (variáveis independentes) e seus resultados formam os ramos da árvore. Cada regra tem início no nó raiz da árvore e caminha até uma de suas folhas (REZENDE et al., 2002).

Os algoritmos que constroem árvores de decisão buscam encontrar aqueles atributos e valores que provêm máxima segregação dos registros de dados, com respeito ao atributo que se quer classificar, a cada nível da árvore.

Normalmente a construção de uma árvore de decisão segue os seguintes passos:

1. Apresenta-se um conjunto de dados e a partir dele é criado o nó inicial (ou nó raiz), com um teste lógico que dividirá a árvore em dois ou mais ramos.
2. A partir da divisão do nó raiz, são gerados outros nós (ou nós internos), sendo que cada nó contém um novo teste lógico, que ramificará novamente a árvore.
3. A divisão dos nós internos continua até que se atinja um nó folha, o qual não irá ramificar mais a árvore.

A repetição deste procedimento caracteriza a recursividade da árvore de decisão (BREIMAN et al., 1984).

As árvores de decisão também podem ser chamadas de árvores de classificação ou de regressão, caso o atributo meta seja categórico ou numérico, respectivamente. Como exemplo simples, a Figura apresenta a árvore de decisão sobre quem compra um computador. O atributo final é binário, sendo classificado como sim ou não (comprar).

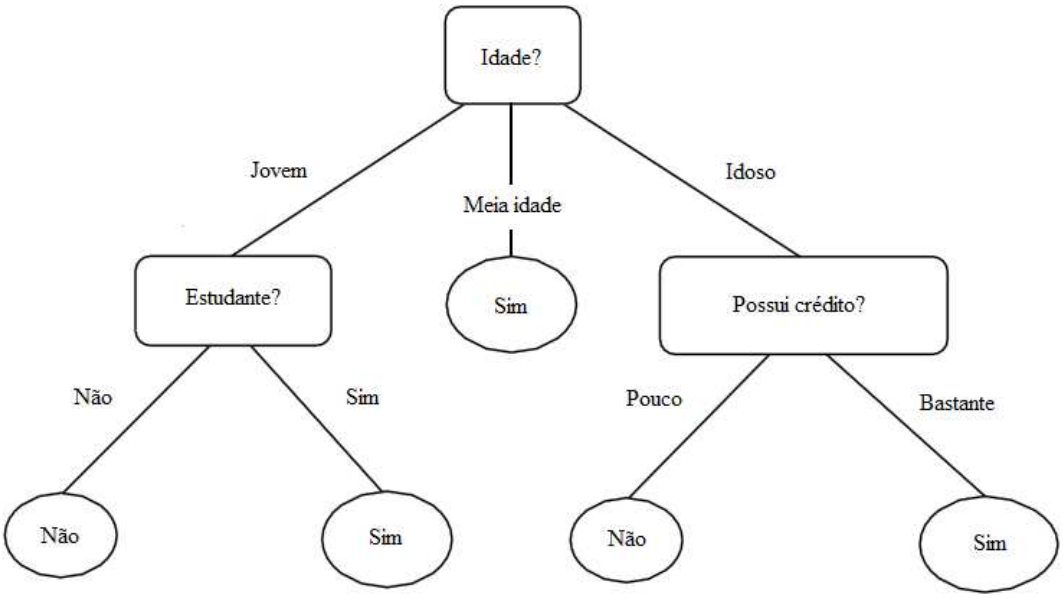

Figura 2: Representação de uma árvore de decisão
Fonte: HAN et al. (2011).

3.2.1.2 Florestas aleatórias

As florestas aleatórias (*Random Forest*) são uma técnica de classificação desenvolvida por Breiman (2001). No algoritmo de árvore de decisão padrão, todo o conjunto de dados é utilizado para formular a árvore, já no algoritmo de florestas aleatórias, o conjunto de dados é dividido aleatoriamente em diversos subconjuntos de tamanho menor. Cada um destes conjuntos é criado por um tipo de amostragem chamado de *bootstrap* (HAN et al., 2011), a qual é do tipo com reposição, ou seja, cada novo conjunto poderá ter alguns registros incluídos mais de uma vez e outros não inclusos nenhuma vez. A amostragem *bootstrap* garante que 1/3 dos exemplos são usados para testar as árvores após sua construção.

A partir de cada subconjunto desenvolvido, uma árvore de decisão é criada. A construção destas árvores ocorre por meio de uma seleção de atributos aleatória dos subconjuntos, os quais são utilizados nos nós de cada uma das árvores desenvolvidas. Uma floresta aleatória é uma coleção dessas árvores de decisão.

Quando a floresta está formada, há um número grande de árvores de decisão a serem testadas e todas contribuem para a classificação do objeto em estudo, por meio de um voto

sobre qual classe o atributo meta deve pertencer. Cada voto tem um certo “peso”, o qual é afetado pela similaridade entre cada árvore, sendo que quanto menor a similaridade entre duas árvores melhor, e pela força que cada árvore tem individualmente, ou seja, quanto mais precisa uma árvore for, melhor será sua nota. O ideal é manter a precisão das árvores sem aumentar sua similaridade (HAN et al., 2011).

O algoritmo é escalar e pode lidar com conjuntos com um grande número de atributos. O uso de subconjuntos e amostragem *bootstrap* faz o algoritmo mais poderoso do que uma simples árvore, apresentando boa taxa de acerto quando testado em diferentes conjuntos de dados (BELLE, 2008). Esta técnica também pode ser considerada uma das mais precisas quando comparada a outras, como redes neurais artificiais e máquinas de vetores suporte (CARUANA et al., 2008). As florestas aleatórias ainda são computacionalmente muito efetivas, além de evitarem sobreajuste (*overfitting*) e serem pouco sensíveis a ruídos (BREIMAN, 2001).

3.2.2 Métodos de seleção de atributos

Seleção de atributos é o processo de refinamento do conjunto de treinamento pela escolha dos atributos mais importantes, sendo que a ideia por trás destes métodos é selecionar os atributos mais importantes do conjunto de dados e ignorar os demais.

A utilização de métodos de seleção de atributos tem duas finalidades. Primeiramente, faz o treinamento e a aplicação de um classificador mais eficiente, diminuindo a quantidade de atributos e tornando o processo de modelagem menos custoso computacionalmente. Em segundo lugar, pode aumentar a precisão de um classificador, eliminando atributos que podem confundi-lo. Tais atributos podem levar a uma generalização incorreta de uma característica acidental do conjunto de treinamento e, consequentemente, a erros de classificação por parte dos modelos (GUYON e ELISSEEFF, 2003).

Existem diversos métodos de seleção de atributos, cada um com uma dada característica. Um deles é o Wrapper, os quais levam em consideração o algoritmo que estará sendo usado sobre o conjunto de dados na fase de modelagem. Wrappers funcionam como uma caixa preta, calculando uma pontuação para um determinado subconjunto. Destes subconjuntos, o mais bem pontuado (que leva a maior taxa de acerto do classificador) é o

selecionado. Wrappers tendem a levar a maior precisão, mas precisam de esforço computacional elevado quando comparados com outros métodos. Este método apresenta ganhos acentuados na taxa de acerto quando o algoritmo de árvores de decisão é utilizado (JOHN e KOHAVI, 1997).

Outro método de seleção de atributos é o CFS (do inglês, *Correlation Feature Selection*). Inicialmente, um conjunto aleatório é escolhido e classificado de acordo com uma medida de correlação com a classe, chamada de mérito. Quanto maior o mérito, mais o conjunto estará relacionado com a classe. O algoritmo busca novos conjuntos com méritos superiores ao primeiro e após cinco testes com subconjuntos de mérito inferior, o conjunto atual é selecionado (HALL, 1999).

Além destes, existem métodos chamados de Infogain e Gainratio, os quais avaliam medidas como o ganho de informação e a taxa de ganho de informação entre atributos e classe (WITTEN et al., 2011).

3.2.3 Medidas de avaliação e desempenho

Diversas medidas de avaliação e desempenho podem ser usadas para analisar o comportamento de um modelo. Nesta seção elas se encontram divididas em três tópicos, o primeiro são medidas de desempenho derivadas de uma matriz que mostra os acertos e erros do modelo, chamada de matriz de confusão, seguido de uma análise gráfica do tipo ROC (*Receiver Operating Characteristics*) e o índice de concordância Kappa.

3.2.3.1 Matriz de confusão

A taxa de acerto e o erro são as medidas de avaliação mais comuns para modelos de classificação (WITTEN et al., 2011). São estimativas dos percentuais de acertos e erros do modelo na predição da classe de novos exemplos. Essas medidas podem ser calculadas a partir da matriz de confusão, que também oferece outros meios efetivos para a avaliação de um classificador (MONARD e BARANAUSKAS, 2002).

Para um problema com duas classes, denominadas classe positiva e classe negativa, a matriz de confusão (Figura 3) indica as quatro possibilidades de acertos e de erros do classificador:

- Verdadeiros positivos (**VP**): quando os exemplos de valor real “SIM” forem preditos como “SIM”.
- Falsos negativos (**FN**): quando os exemplos de valor real “SIM” forem preditos como “NÃO”.
- Verdadeiros negativos (**VN**): quando os exemplos de valor real “NÃO” forem preditos como “NÃO”.
- Falsos positivos (**FP**): quando os exemplos de valor real “NÃO” forem preditos como “SIM”.

	Predição: SIM	Predição: NÃO
Valor real: SIM	VP	FN
Valor real: NÃO	FP	VN

Figura 3: Exemplo de uma matriz de confusão.

Outras medidas de avaliação também podem ser derivadas da matriz de confusão (HAN et al., 2011): sensitividade, especificidade, confiabilidade positiva, e confiabilidade negativa.

A sensitividade *(Recall ou TPR – True Positive Rate)* é a proporção de exemplos positivos que foram classificados corretamente, já a especificidade *(TNR – True Negative Rate)* é a proporção de exemplos negativos que foram classificados corretamente.

A confiabilidade positiva *(Precision ou PPV – Positive Predicted Value)* é a proporção de exemplos positivos classificados corretamente dentre todos os exemplos classificados como positivos. A confiabilidade negativa *(NPV – Negative Predicted Value)* é a proporção de exemplos negativos classificados corretamente dentre todos os exemplos que foram classificados como negativos.

As equações de cálculo das medidas de avaliação são as seguintes:

$$Taxa.de.acerto = \frac{VP + VN}{n} \quad (1)$$

$$Erro = \frac{FP + FN}{n} \quad (2)$$

$$Sensitividade = \frac{VP}{VP + FN} \quad (3)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (4)$$

$$Confiabilidade.Positiva = \frac{VP}{VP + FP} \quad (5)$$

$$Confiabilidade.Negativa = \frac{VN}{VN + FN} \quad (6)$$

Onde n é o número total de exemplos no conjunto.

As medidas de avaliação de um modelo podem ser geradas por diversas técnicas de amostragem, uma delas chama-se validação cruzada. Trata-se de um método que está baseado na divisão do conjunto de dados em N partições, mutuamente exclusivas, de tamanho igual. O modelo é gerado a partir de N-1 partes, e testado na parte que foi removida do conjunto de treinamento. Este procedimento é repetido N vezes, até que todas as partes tenham sido usadas para teste do modelo. A validação cruzada realizada com 10 partições permite obter as melhores medidas de avaliação sobre um modelo (WITTEN et al., 2011).

3.2.3.2 O índice Kappa

O índice Kappa foi introduzido por Cohen (1960). Este índice estatístico é uma medida de concordância usada em escalas nominais. Ele fornece subsídios sobre quanto as classificações são diferentes daquelas esperadas (ao acaso). No caso mais específico de mineração de dados, ele mede a correlação entre os valores preditos e os observados, corrigindo uma correlação que ocorre ao acaso (WITTEN et al., 2011).

Para uma matriz de confusão (Figura 3), o índice Kappa pode ser calculado por meio da determinação da probabilidade esperada, chance de ocorrer correlação ao acaso, e da probabilidade observada, correlação que realmente ocorreu. Isto pode ser feito seguindo as equações (8), (9) e (10):

Cálculo da Probabilidade Esperada (Pe):

$$Pe = \frac{(VP + FN) * (VP + FP) + (VN + FP) * (VN + FN)}{n^2} \quad (8)$$

Cálculo da Probabilidade Observada (Po):

$$Po = \frac{(VP + VN)}{n} \quad (9)$$

Cálculo do índice Kappa (IK):

$$IK = \frac{(Po - Pe)}{(1 - Pe)} \quad (10)$$

Os valores do índice Kappa podem chegar até 1, onde valores intermediários representam diversos tipos de classificação quanto à correlação, ou ao comportamento de um modelo, demonstrados na Tabela 2.

Tabela 2: Índices de avaliação Kappa.
Fonte: Adaptado de LANDIS e KOCK (1977)

Valor de IK	Classificação do modelo
Menor de 0,40	Ruim
0,41 – 0,60	Regular
0,61 – 0,80	Bom
Maior de 0,81	Excelente

3.3 A cultura do café e o mapeamento de áreas cafeeiras

3.3.1 A cultura do café

O cafeeiro é uma planta da Família *Rubiaceae* e pode ser classificado em dois gêneros, *Coffea* e *Psilanthus*. Todas as espécies do gênero *Coffea* são nativas de regiões tropicais da África e ilhas do Oceano Índico, já as espécies do gênero *Psilanthus* são originárias da África, Ásia e Oceania (CROS et al., 1998). O gênero *Coffea* compreende duas espécies: *Coffea arabica* Linnaeus (café arábica) e *Coffea canephora* Pierre. (café robusta). Estas espécies são responsáveis por praticamente todo o café comercializado mundialmente (FAZUOLI et al., 2002; MATIELLO et al., 2002).

A espécie *Coffea arabica* é nativa de uma região da África que compreende o sudoeste da Etiópia, sudeste do Sudão e norte do Quênia, regiões com altitude variando entre 1000 e 2000 metros. Já o *Coffea canephora* possui uma distribuição geográfica muito mais ampla, ocorrendo em grande parte do continente africano.

O café robusta apresenta um menor custo de produção, uma alta produtividade e um maior rendimento quando comparado ao café arábica, entretanto, o café arábica é mais aceito no mercado por possuir melhor sabor e aroma (MENDES et al., 2001).

O café foi introduzido no Brasil em 1727, vindo de plantações da Guiana Francesa. A frutificação do café ocorre, em média, cerca de dois anos após o plantio, dependendo dos tratos culturais utilizados. A brota do café depende muito do clima e da altitude do cultivo. Sua flor dá origem a um fruto de cor vermelha ou amarela, com aproximadamente 10 a 15 milímetros de diâmetro.

O Brasil tem cerca de 2,2 milhões de hectares plantados com café e produziu, em 2012, 55,9 milhões de sacas. O café é cultivado em 12 estados brasileiros, sendo que os três maiores produtores são Minas Gerais, com cerca de 51% da produção nacional, seguido do Espírito Santo, com 27%, e São Paulo, com 8% (MINISTÉRIO DA AGRICULTURA, 2013).

No âmbito mundial, o Brasil é o maior produtor e exportador de café em grão, tendo exportado cerca de 33,6 milhões de sacas (60% de sua produção) para seus principais importadores, Estados Unidos e União Européia, gerando ganhos de aproximadamente US$ 6

bilhões (USDA, 2013). A importância econômica do café é ainda mais exaltada por este ser, junto com o açúcar, a principal commoditie agrícola do Brasil, além de ser a 3ª commoditie mais exportada pelo país, ficando atrás apenas de petróleo bruto (2ª) e minério de ferro (1ª). (MINISTÉRIO DO DESENVOLVIMENTO, 2013)

3.3.2 Classificação de áreas cafeeiras

Os primeiros estudos relacionados ao mapeamento de áreas cafeeiras foram realizados por Dallemand (1987) e Batista et al. (1990). O estado de Minas Gerais teve sua área cafeeira mapeada por Moreira et al. (2007). Este trabalho visou mapear todo o estado e identificar áreas cafeeiras e não cafeeiras para imagens do satélite Landsat 5 do ano de 2006. Os resultados mostraram que a região sul de minas é responsável por cerca de 50% da área de café plantada do estado, ressaltando ainda mais sua importância no contexto nacional. Dentre outras conclusões obtidas pelos autores está a dificuldade de identificação de áreas cafeeiras com áreas de cerrado e eucalipto.

Adami et al. (2009), avaliaram a exatidão do mapeamento realizado por Moreira el al. (2007), quanto a identificação de áreas cafeeiras. Os autores concluíram que 95% do mapeamento foi feito corretamente, mais precisamente para a região sul e sudeste de minas, essa taxa de acerto foi de 99%. Tal exatidão foi atribuída à característica de cultivo da região. Estes resultados mostram o potencial que as geotecnologias têm sobre a identificação de áreas cafeeiras na região.

Vieira et al. (2006) estudaram a relação entre os parâmetros da cultura cafeeira e a resposta espectral em imagens do sensor TM/Landsat, em áreas relevantes para a cafeicultura em Minas Gerais. Os autores concluíram que a resposta espectral do café se assemelha muito com a da vegetação natural. Como as curvas espectrais da mata e do café são muito parecidas, é comum e esperada uma confusão entre essas duas classes.

Em regiões menores também encontram-se aplicações de sensoriamento remoto aplicada à áreas cafeeiras. Como exemplo, Moreira et al. (2008) avaliaram o uso de imagens

dos satélites Landsat 5 e Quickbird para mapeamento de áreas cafeeiras em diversas microrregiões da região sul e sudeste de minas gerais. Eles concluíram que imagens com uma resolução espacial maior produzem, geralmente, melhores resultados na classificação. A mesma conclusão foi obtida por Ramirez et al. (2006), entretanto comparando imagens do satélite Landsat 5 e Ikonos 2, chegando ao resultado que as imagens de alta resolução (Ikonos 2) identificaram 1,5 vez mais talhões de café do que imagens de média resolução (Landsat). Já Santos et al. (2011) mapearam áreas cafeeiras em microrregiões do estado do Espírito Santo com aerofotografias, mostrando que estas imagens, de altíssima resolução espacial, são capazes de identificar de áreas cafeeira com ótima exatidão, permitindo até determinar qual sistema de cultivo está sendo utilizado.

Trabalhos relacionando a área de mineração de dados e identificação de áreas cafeeiras já foram realizados. Marques (2003) realizou uma classificação supervisionada, pixel a pixel, utilizando diversos algoritmos, dentre eles o Maxver (máxima verossimilhança) para uma imagem da região cafeeira de Machado (MG). O mapa elaborado pelo algoritmo Maxver obteve um índice Kappa de 0,39. Já Bernardes et al. (2007), utilizando imagens Landsat, avaliaram classificadores automáticos (pelos algoritmos de Maxver, Isoseg e Battacharya) para áreas cafeeiras no município de Patrocínio-MG. Os classificadores obtiveram melhor índice Kappa de 0,31, para o Isoseg. Em ambos os casos os autores avaliaram que em regiões predominantemente planas a classificação automática apresenta melhores resultados, a medida que em regiões com relevo movimentado a classificação torna-se mais complexa, devido a diferentes padrões de sombreamento nas áreas cafeeiras.

Ainda justificando a influência do relevo na classificação automática de imagens orbitais, Machado (2002) encontrou índices de acerto extremamente baixos para classificação supervisionada de imagens visando a identificação da cultura cafeeira na região de Manhuaçu (MG). Os valores encontrados variaram de 25 a 42% e o autor explica que problemas relacionados à similaridade espectral, à topografia acidentada da região, com o consequente sombreamento das imagens, e à fragmentação das lavouras de café, localizadas em áreas contíguas a matas, foram os principais causadores de baixo desempenho.

Andrade et al. (2011), avaliaram redes neurais para o mapeamento de áreas cafeeiras na região de Três pontas (MG), utilizando de imagens do satélite Landsat 5. O resultado da classificação obteve um índice Kappa de 0,67, sendo que o principal erro apontada pelos autores foi a ambiguidade de classificação entre áreas de café e mata, tendo em vista os padrões espectrais muito próximos destas duas classes. Posteriormente, Andrade et al. (2013), realizaram um procedimento análogo para a região de Machado (MG). Para este estudo, os autores separaram as regiões de relevo acentuado das de relevo plano, para posterior classificação. Nos locais de relevo mais movimentado o índice Kappa foi de 0,55, a medida que para locais planos foi de 0,60. Neste estudo foram aplicados máscaras de drenagem e área urbana, a fim de diminuir a variabilidade dos alvos classificados.

4 Material e Métodos

A metodologia usada para desenvolver o processo de descoberta de conhecimento (*KDD – Knowledge Discovery in Databases*) foi a CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (CHAPMAN et al., 2000). Esta divide o ciclo de vida de um projeto em 6 fases: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição (Figura 4). A metodologia é cíclica, sendo que a sequência lógica entre as fases não é rígida, sendo comum, e quase sempre necessário, voltar e avançar entre diferentes fases. O resultado de uma fase realizada anteriormente é fundamental para determinar qual a próxima fase que deverá ser executada na sequência.

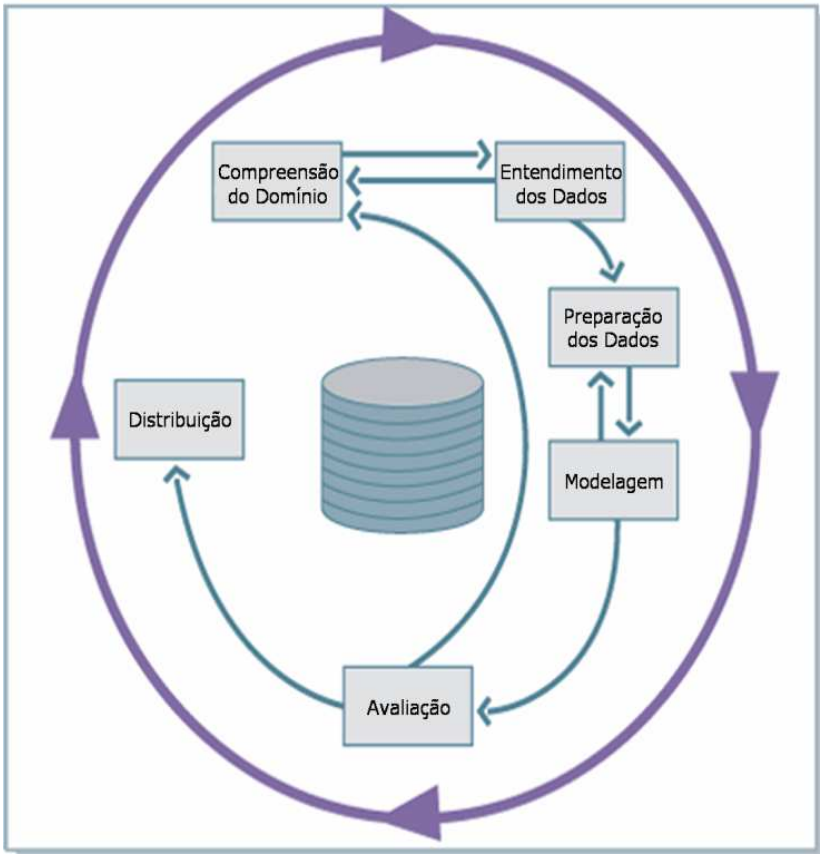

Figura 4: Fases do modelo de processo CRISP-DM.
Fonte: Adaptado de CHAPMAN et al. (2000).

Fases do processo KDD:

Compreensão do domínio: Esta fase inicial visa entender os objetivos e requisitos do projeto, pela perspectiva do domínio de aplicação, e depois converter esse conhecimento na definição do problema e em um plano projetado para atingir os objetivos.

Entendimento dos dados: A fase de entendimento dos dados começa com a coleta inicial de dados e continua com atividades para se familiarizar com os dados, identificar problemas de qualidade nesses dados e buscar as primeiras compreensões (*insights*) a partir deles.

Preparação dos dados: Fase que abrange todas as atividades necessárias para construir, a partir dos dados iniciais brutos, o conjunto de dados final para a modelagem. Atividades como transformação de dados e geração dos atributos do conjunto de dados são realizadas nesta fase.

Modelagem: Várias técnicas de mineração de dados são selecionadas e aplicadas nesta fase, além de seus parâmetros serem calibrados para valores ótimos. Existem várias técnicas que podem ser utilizadas em um mesmo problema, algumas têm requisitos específicos para o conjunto de dados, então voltar para a fase de preparação de dados é normalmente necessário.

Avaliação: Esta fase inicia-se com um ou mais modelos construídos. Estes modelos apresentam, aparentemente, boa qualidade na perspectiva da análise de dados. Antes de proceder com sua distribuição, é importante avaliar cada modelo de forma mais completa e rever os passos executados na sua construção, para se ter a certeza de que atendem aos objetivos traçados

Distribuição: Esta fase finaliza o processo colocando os modelos gerados disponíveis para uso, mesmo que seja com a intenção de aumentar o conhecimento obtido pela extração de padrões do conjunto de dados.

4.1 Entendimento dos dados

4.1.1 O conjunto de dados

Os dados utilizados foram coletados junto ao pesquisador Maurício Alves Moreira do INPE e seu aluno de mestrado Renan Marujo por contato pessoal em 2014. Os dados referem-se a classificação manual de café para a região sul de minas gerais utilizando uma imagem do LANDSAT-8 para o ano de 2014 (08/03/2014). O município selecionado para o estudo foi Machado, por ser um dos principais produtores do estado e já ter sido alvo de outros estudos relacionados a classificação automática de café (Figura 5). Ele está localizado na região sul de minas gerais, latitude sul de 21º 40’ 30’’, longitude oeste de 45º 55’ 12’’ com altitude média de 835 m e uma área de 585,3 Km^2.

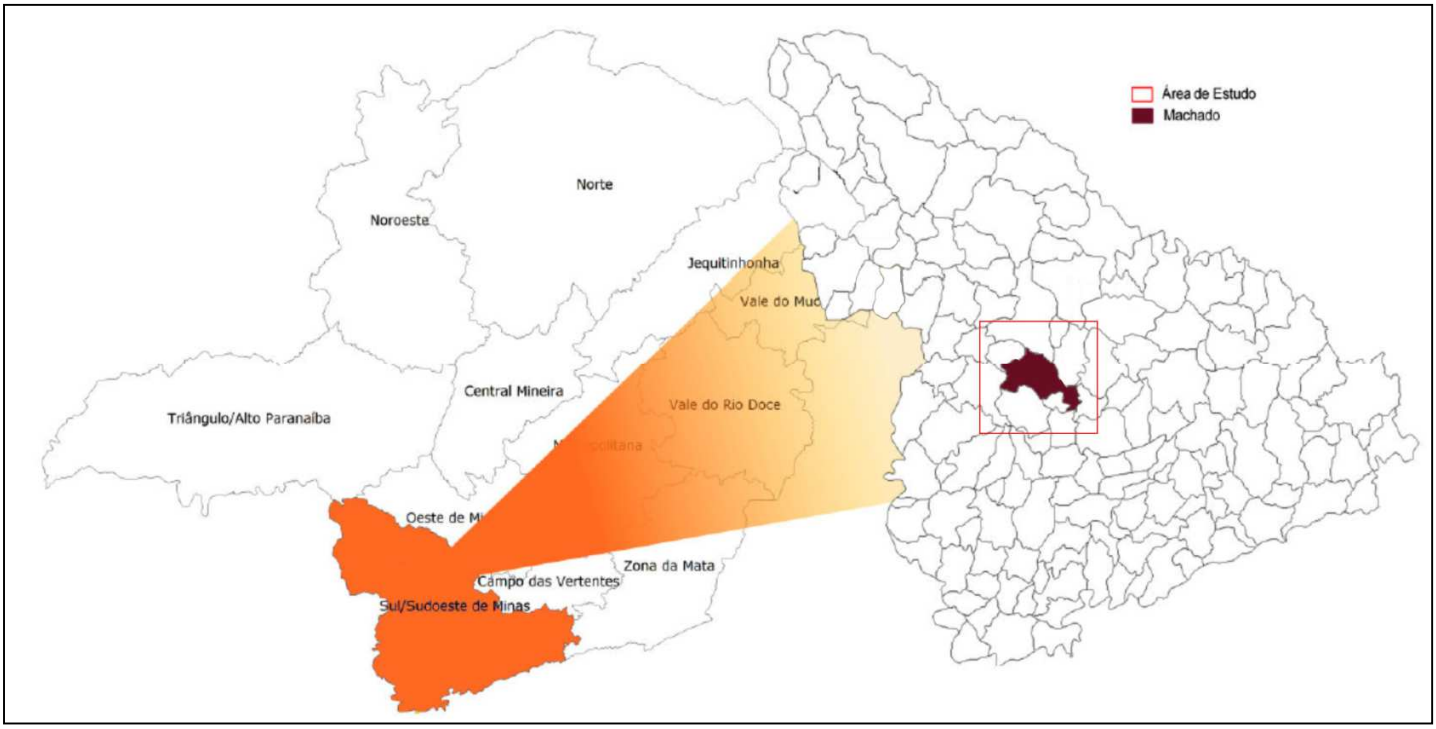

Figura 5: Área de estudo – município de Machado/MG.

Arquivos obtidos:

O conjunto de dados obtido foi composto por três tipos de arquivos.

- Uma imagem com 11 bandas do LANDSAT-8 para a região sul de minas gerais.
- Um Shapefile contendo os estados de minas gerais.
- Uma imagem temática com classificação de áreas de café para a mesma região da imagem.
- Um Shapefile contendo as principais máscaras de drenagem do estado de minas.

4.2 Preparação dos dados

4.2.1 Transformação dos dados

A transformação dos dados foi um processo que visou modificar os dados obtidos em dados que pudessem ser utilizados no processo de mineração de dados. Este procedimento foi o mais trabalhoso e foi dividido em diversos Passos.

Passo 1 – Reunião dos dados brutos em um sistema de informação geográfica

O primeiro passo da preparação dos dados teve inicio com a junção de todos os arquivos obtidos em um software de sistema de informação geográfica (SIG). O software utilizado foi o spring, versão 5.1.8 (CÂMARA et al., 1996).

Foram inseridas as imagens relacionadas as 11 bandas do LANDSAT-8, a imagem temática contendo áreas classificadas de café, o arquivo shapefile com os limites dos municípios da cidade e o shapefile com as principais redes de drenagem, como pode ser visto nas Figura 6 a 8. Também foram geradas imagens sintéticas em diversas composições de cores falsas e infravermelho, Figuras 9 a 11.

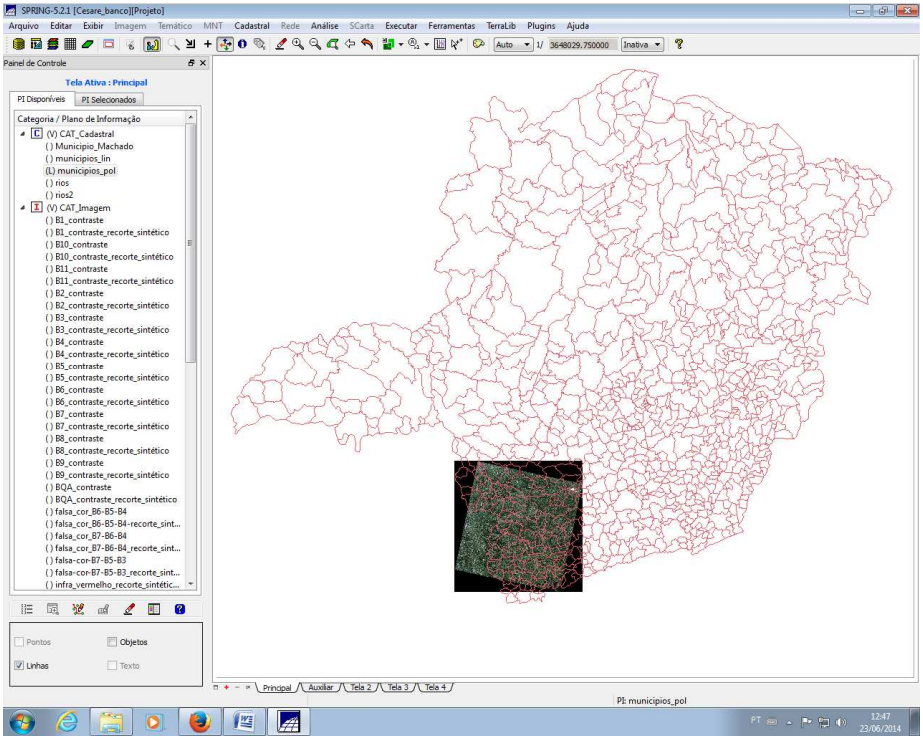

Figura 6: Visualização na área de estudo para o estado de Minas Gerais.

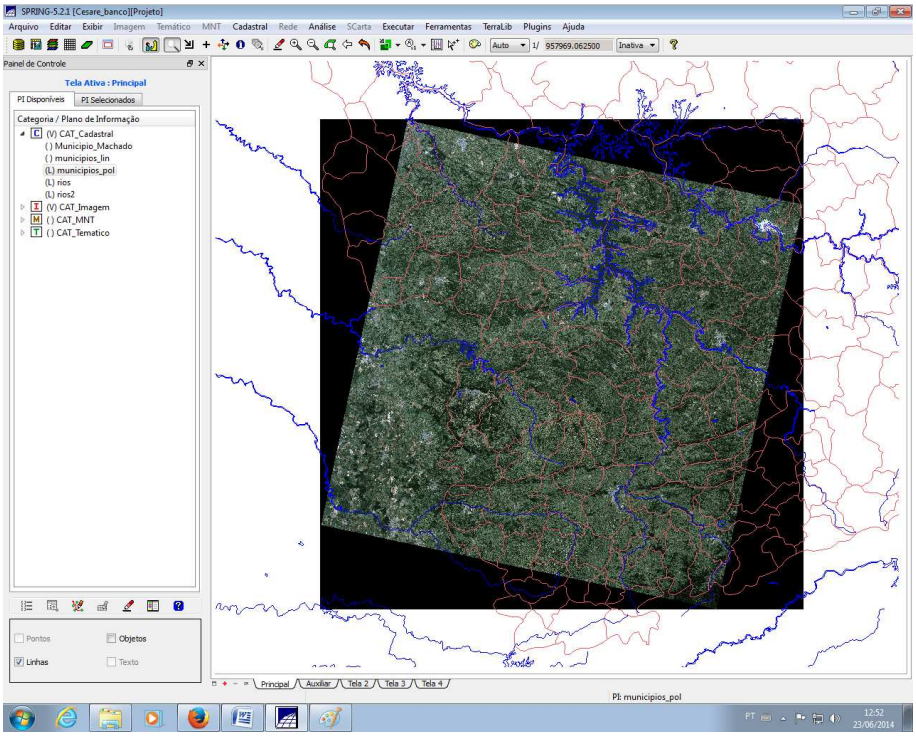

Figura 7: Visualização na área de estudo para a imagem do LANDSAT8 com shape dos rios e municípios de Minas Gerais.

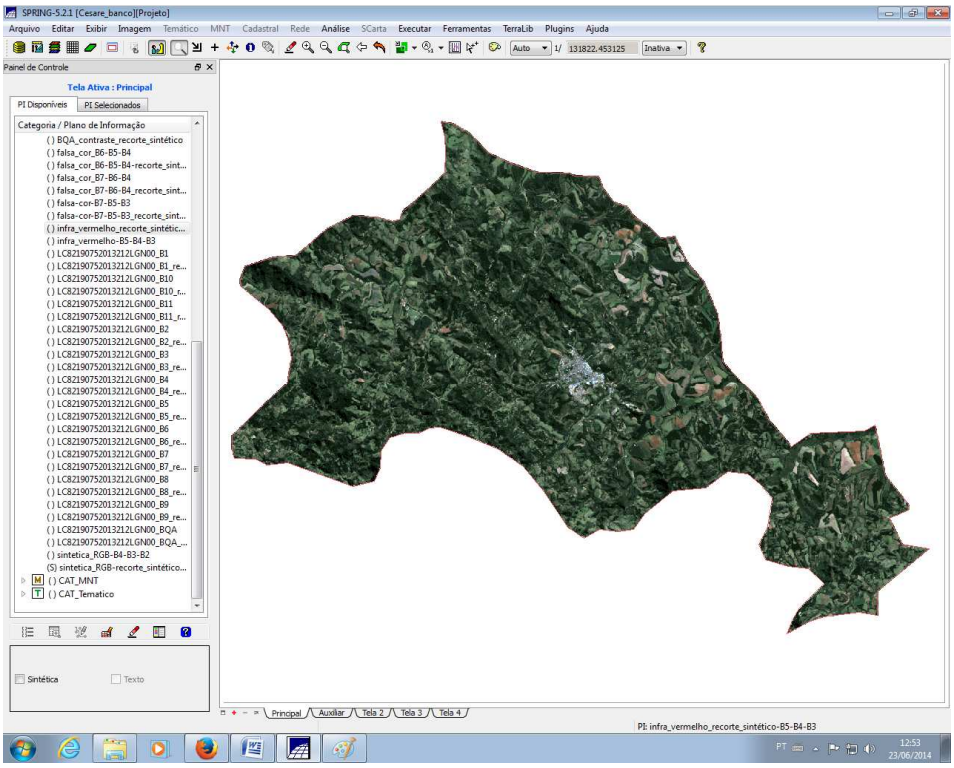

Figura 9: Composição sintética para as bandas 2, 3 e 4 para a área de estudo.

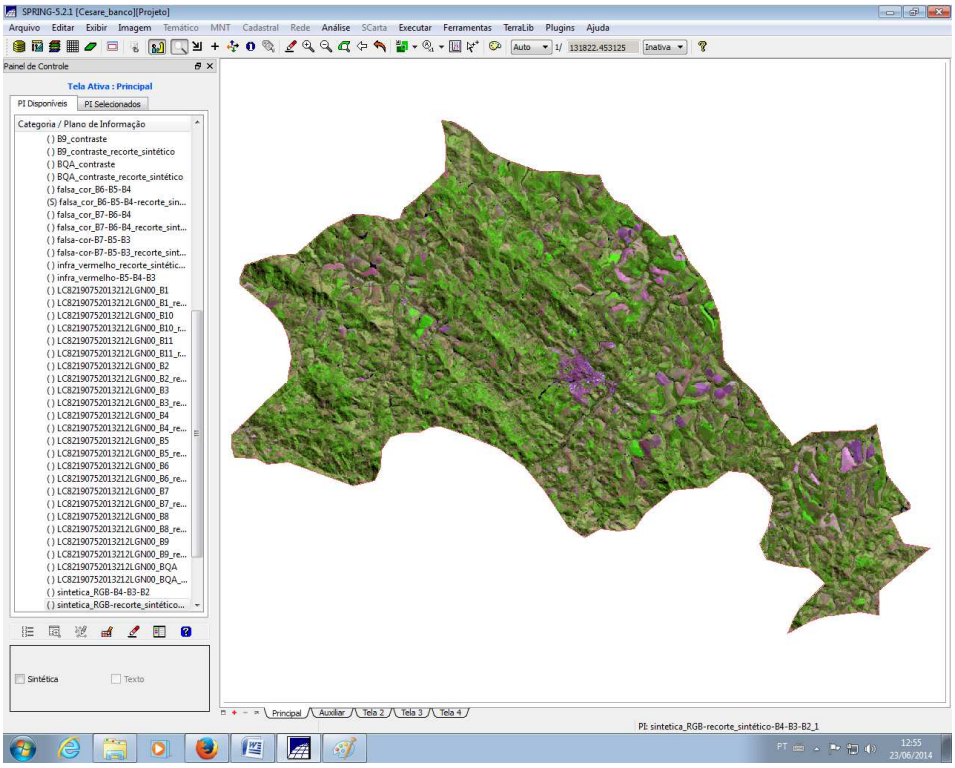

Figura 10: Composição sintética para as bandas 6, 5 e 4 para a área de estudo.

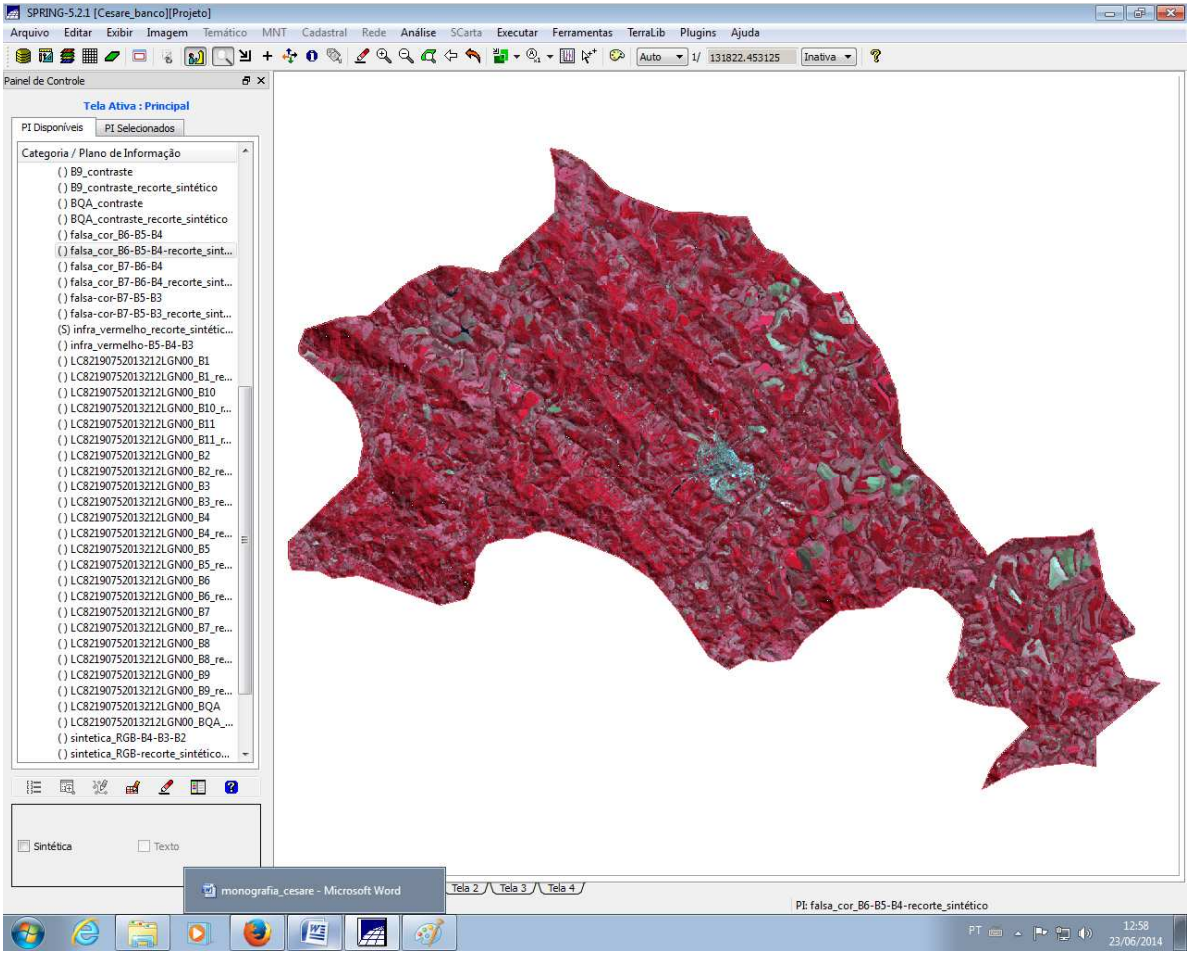

Figura 11: Composição sintética em infravermelho (bandas 5, 4 e 3) para a área de estudo.

Passo 2 - Criação de atributos de nível de cinza

O segundo passo da preparação dos dados foi criar os atributos de nível de cinza para cada uma das bandas do LANDSAT-8. Inicialmente foi feito um recorte da área de estudos para o município de machado, utilizando seu formato do arquivo shapefile (Figuras 12).

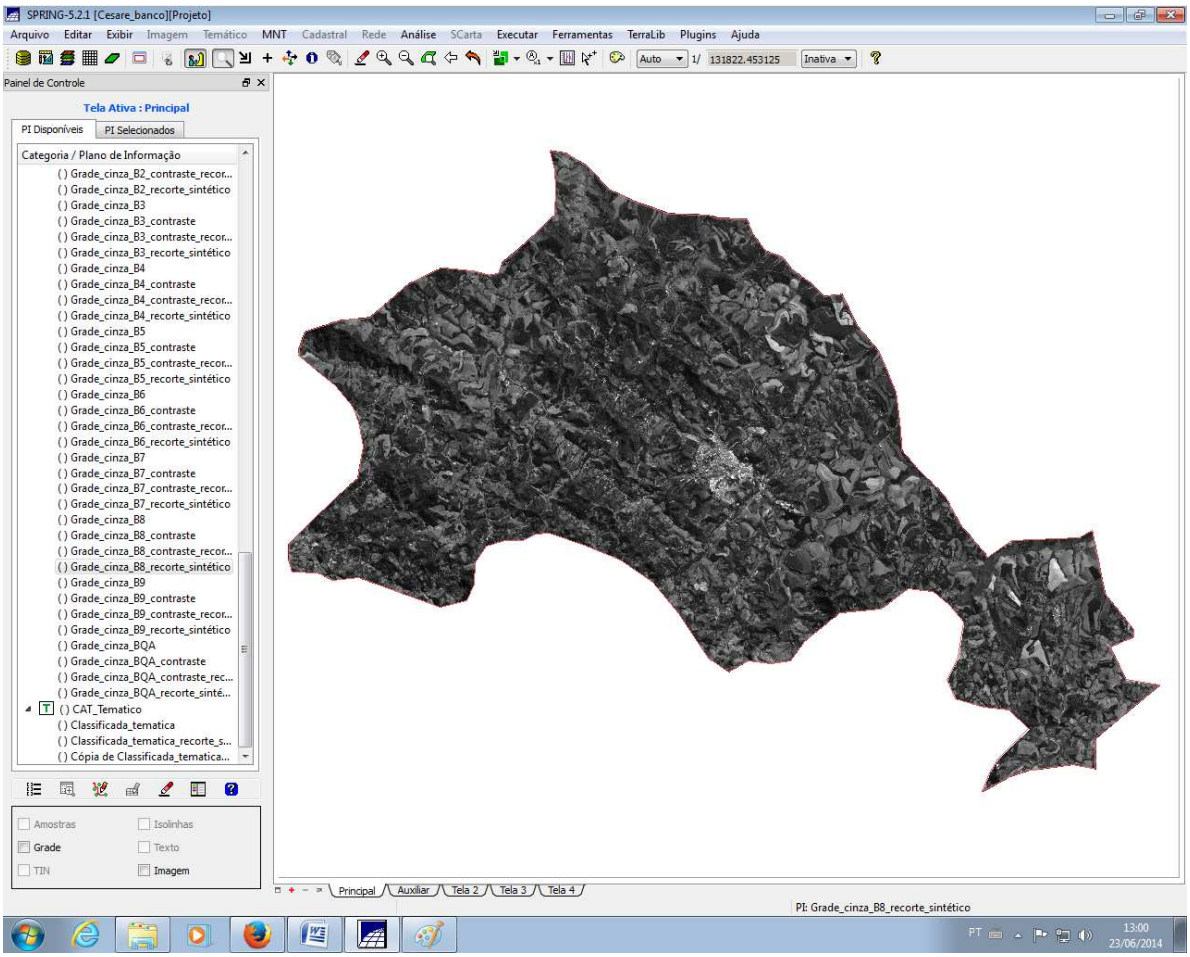

Figura 12: Imagem de cinza para a banda pancromática para a área de estudo.

Com o recorte da área do município de Machado foi realizada uma modelagem numérica de terreno (MNT), afim de se obter o nível de cinza dos pixels nesta região (Figura 13 e 14). Este procedimento foi repetido para cada uma das bandas.

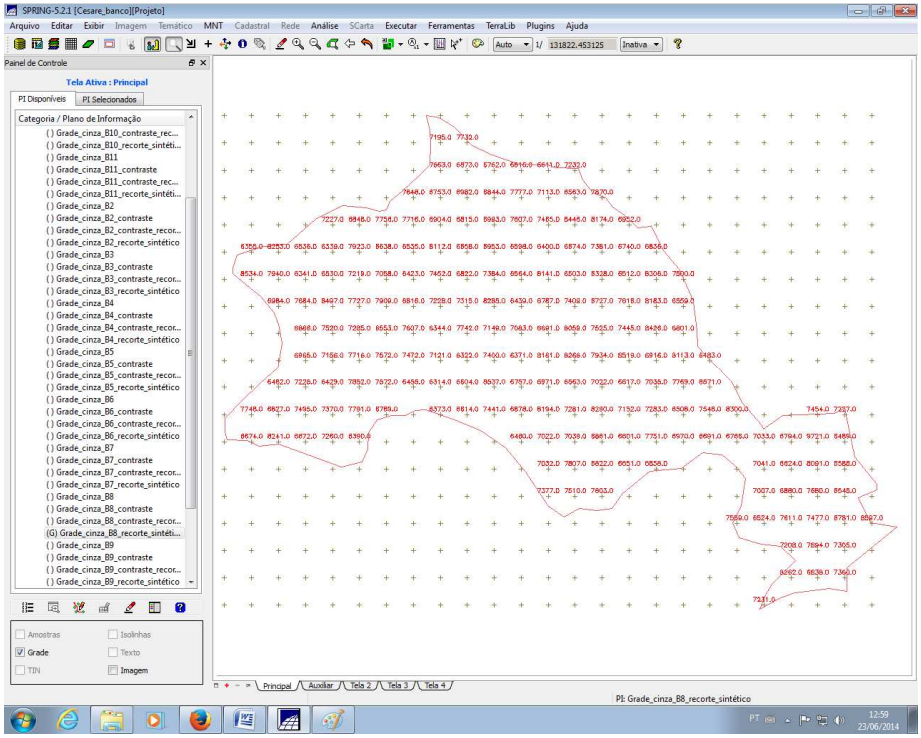

Figura 13: Geração do MNT em nível de cinza para a área de estudo.

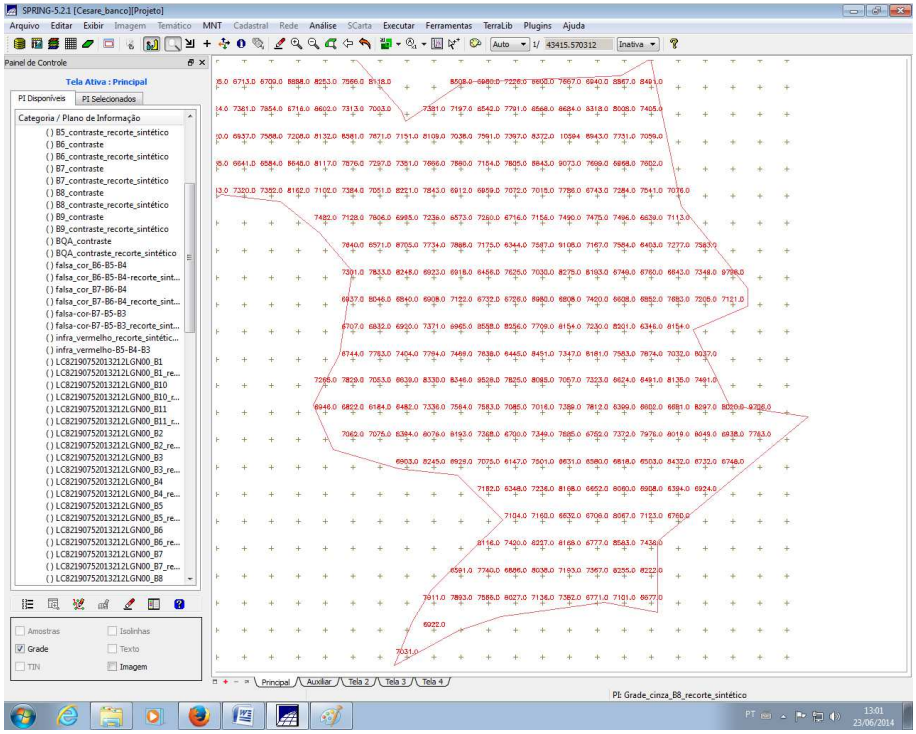

Figura 14: Geração do MNT em nível de cinza para a área de estudo com zoom para um local do estado.

Passo 3 - Extração dos atributos para modelagem

O terceiro passo da preparação dos dados foi extrair os atributos do software SPRING e criar um arquivo que pudesse ser utilizado no software WEKA, versão 3.6.11 (HALL et al., 2009). Este software é o responsável por realizar o processo de mineração de dados e extração de padrões embutidos nos dados.

Este passo contou com a ajuda do pesquisador do INPE Thales Sehn Korting. Foram adaptados scripts na linguagem computacional C++ para a extração destes dados do software SIG, resultando em um arquivo do tipo . CSV, com todos os atributos necessários para realizar a mineração de dados.

4.2.2 Atributos do conjunto de dados

Atributo Meta:

O atributo meta ou a variável dependente foi a classe nomeada café ou não café. Ela foi obtida de acordo com a classificação manual de café realizada por Moreira para cada pixel dentro do área de estudo. Na imagem classificada também foram mapeadas classes relacionadas ao perímetro urbano de Machado e também pequenas áreas classificadas como café podado. No processo de poda o café sofre, normalmente, uma redução de sua área a fim de aumentar a sua produtividade, entretanto a poda pode ser de diversas formas, passando por procedimentos como recepa, decote, esqueletamento e desponte (PROCAFÉ, 2007). Consequentemente pode ocorrer uma mudança no comportamento espectral nestas áreas assim, a classe de café podado foi excluída do conjunto de treinamento. O mesmo foi realizado para a classe urbana, seguindo procedimento análogo a Andrade et al., (2013).

Atributos preditivos:

Os atributos preditivos, ou variáveis independentes, partiram do valor de cinza de cada pixel. Foram consideradas todas as bandas, mesmos aquelas que não são específicas para estudos relacionados a vegetação. A tabela 3 apresenta os atributos do conjunto de dados.

Tabela 3: Atributos presentes no conjunto de dados.

Nome	Tipo	Significado
BANDA 1	numérico	Nível de cinza dos pixels na banda 1.
BANDA 2	numérico	Nível de cinza dos pixels na banda 2.
BANDA 3	numérico	Nível de cinza dos pixels na banda 3.
BANDA 4	numérico	Nível de cinza dos pixels na banda 4.
BANDA 5	numérico	Nível de cinza dos pixels na banda 5.
BANDA 6	numérico	Nível de cinza dos pixels na banda 6.
BANDA 7	numérico	Nível de cinza dos pixels na banda 7.
BANDA 8	numérico	Nível de cinza dos pixels na banda 8.
BANDA 9	numérico	Nível de cinza dos pixels na banda 9.
BANDA 10	numérico	Nível de cinza dos pixels na banda 10.
BANDA 11	numérico	Nível de cinza dos pixels na banda 11.
CLASSE	nominal	Classe do pixel: "café" ou "não-café"

O conjunto de dados preparado totalizou 645.785 registros.

4.3 Modelagem

A fase de modelagem foi dividida em duas grandes etapas: a primeira foi a de pré-indução e a segunda de indução. Na primeira etapa, o conjunto de dados foi separado em uma amostra e, em seguida, foram aplicadas as técnicas de seleção de atributos. Já na segunda fase, ocorreu a indução dos modelos utilizando as técnicas de mineração de dados. A modelagem foi realizada em um processo cíclico junto com a fase de avaliação dos modelos.

4.3.1 Etapa de pré-indução

4.3.1.1 Amostragem

A partir do conjunto de dados brutos foi realizado um procedimento de amostragem sem reposição. O processo de amostragem permite que os algoritmos de mineração tratem enormes bases de dados pela redução do número de casos avaliados sem que haja perdas significativas na qualidade das predições (WITTEN et al., 2011).

A amostragem foi feita pelo software WEKA com um valor de conservação de 0,25, ou seja, 25% dos registros originais no conjunto amostrado pertencem ao original. Além disso a amostragem foi realizada mantendo o equilíbrio das classes em 50% cada uma. Classes balanceadas tendem a melhores resultados do que uma distribuição diferente (WEISS e PROVOST, 2001). O conjunto reamostrado totalizou 161.446 registros.

4.3.1.2 Métodos de seleção de atributos

A partir do conjunto de dados preparado e balanceado, foram utilizados os métodos de seleção de atributos. Estes métodos têm como finalidade reduzir o número de atributos do conjunto, melhorando o desempenho dos modelos gerados e proporcionando um ganho computacional.

Foram aplicados métodos em que um algoritmo de seleção, já implementado, é utilizado para filtrar o conjunto de dados. Eles foram:

- Wrapper (WRP)
- Correlation Feature Selection (CFS)
- InfoGain (IG)
- GainRatio (GR)

O wrapper e o CFS eliminam os atributos pouco relevantes do conjunto de dados, diferentemente dos demais métodos, os quais são responsáveis por ranquear os atributos de acordo com sua contribuição. A partir desse ranking, foram feitas diversas escolhas de atributos manualmente até se chegar ao melhor conjunto. O ranking dos atributos é o mesmo independentemente da técnica de modelagem, entretanto, a seleção manual mudou de técnica para técnica.

4.3.2 Fase de indução

4.3.2.1 Geração de novos modelos

A fase de indução utilizou duas técnicas de modelagem para gerar os modelos:

- Árvores de decisão
- Florestas aleatórias

Os parâmetros de cada algoritmo foram calibrados visando melhor taxa de acerto na classificação. As árvores de decisão foram geradas por meio do classificador nomeado “J48” no WEKA. Uma opção utilizada foi o número mínimo de objetos por folha, a qual gera uma árvore onde cada folha teve uma quantidade mínima de registros. Quanto menos objetos por folha, maior ficava a árvore e melhor era sua taxa de acerto. Porém, um número muito pequeno de registros por folha poderia gerar um sobreajuste no modelo, pois ele acertaria casos específicos ao conjunto de treinamento.

Já as Random Forest foram avaliadas em de dois parâmetros de calibração. Estes dois parâmetros foram a profundidade das árvores e o número de atributos aleatórios utilizados nas árvores. A quantidade de árvores geradas em cada floresta foi definida em 25, valor inferior ao recomendado por Breiman (2001). Entretanto, notou-se que a partir deste valor, não houve ganho na taxa de acerto, apenas um maior custo computacional para gerar os modelos.

5 Desenvolvimento de modelos de classificação

Este capítulo trata da construção dos modelos de classificação automática para áreas cafeeiras de Machado (MG). Foram desenvolvidos mais de 100 modelos, seguindo o proposto na seção anterior.

5.1 Modelos avaliados no conjunto de treinamento

Esta seção apresenta os modelos desenvolvidos e avaliados junto ao conjunto de treinamento. Nesta etapa foram avaliados modelos em árvores de decisão e Random Forest. Foi realizada uma avaliação inicial destes sobre o conjunto de treinamento, o que se trata de uma estimativa otimista, pois todos os registros classificados foram utilizados para treinar os modelos.

Tabela 4: Árvores de decisão avaliadas no conjunto de treinamento.

	Árvores de decisão variando-se o número de registros por folha					
	30000	15000	7500	1500	750	250
Taxa de acerto	71,75%	72,15%	73,44%	75,76%	76,53%	78,40%
Índice Kappa	0,43	0,44	0,47	0,51	0,53	0,56
Número de folhas	3	4	7	27	39	118

As árvores de decisão não demonstraram um resultado satisfatório quando utilizadas no conjunto de treinamento. O melhor índice Kappa foi de 0,56 com uma taxa de acerto de 78,40%. Entretanto, a árvore que apresentou 118 folhas, o que a deixa praticamente impossível de interpretar.

Algo que chamou a atenção nas árvores de decisão foi a presença das bandas 3, 5 e 8 sempre na parte inicial da árvores. Isto indica que o algoritmo encontrou estes atributos como principais para classificação no conjunto de dados (quanto mais acima está um atributo, mais entropia ele possui, e, conseqüentemente, mais eficiência em classificar os registros). Estas bandas são intimamente relacionadas a classificação da vegetação, sendo a banda 3 (Verde)

responsável pela reflectância da vegetação sadia e a banda 5 (infravermelho próximo) para a identificação de culturas agrícolas e diferenciação delas com o solo.

Quando comparadas as árvores de decisão, as Random Forest obtiveram um desempenho muito superior. Este desempenho foi praticamente perfeito em alguns casos, mas como mencionado anteriormente, ele esta sendo avaliando em cima do conjunto de treinamento.

Tabela 5: Random Forest avaliadas no conjunto de treinamento.

	Random Forest variando-se o número de árvores na floresta			
	5	10	25	100
Taxa de acerto	97,20%	98,62%	99,78%	99,99%
Índice Kappa	0,94	0,97	0,99	1,00

A quantidade de árvores influenciou em pouco a taxa de acerto do algoritmo. Pela tabela 5, pode-se notar que 25 árvores na floresta já apresentam praticamente um desempenho perfeito, apenas 347 registros foram classificados incorretamente neste caso. Para o caso mais otimista, com 100 árvores na floresta, apenas 4 registros foram classificados incorretamente como "não-café", a medida que a classificação correta seria de áreas de café. A variação dos outros parâmetros internos não melhorou na classificação das florestas.

A partir deste ponto pode-se notar que o algoritmo mais eficiente para este tipo de classificação foram as random forest. Entretanto, deve-se avaliar o desempenho de ambos em um conjunto de teste, composto por todas os registros referentes ao município de machado, com exceção das classes de café podado e áreas urbanas.

5.2 Modelos avaliados para o conjunto de teste

A fim de saber com uma maior confiabilidade qual modelo é o responsável para inferir sobre áreas cafeeiras no município de machado, os modelos anteriores foram avaliados para um conjunto de teste, ou seja, para dados que não foram os mesmos utilizados no seu treinamento. Este resultado apresenta uma perspectiva mais real do que avaliar um modelo apenas com o conjunto de treinamento, pois ele está avaliando-o em uma situação real, que

não fez parte de seu treinamento. O conjunto de teste conteve todos os pixels para a região de machado, totalizando 645.785 registros.

Tabela 6: Árvores de decisão avaliadas no conjunto de teste.

	Árvores de decisão variando-se o número de registros por folha					
	30000	15000	7500	1500	750	250
Taxa de acerto	68,73%	75,96%	76,10%	75,93%	77,47%	76,90%
Índice Kappa	0,30	0,36	0,38	0,40	0,42	0,42
Número de folhas	3	4	7	27	39	118

Ao serem avaliadas no conjunto de treinamento pode-se notar uma redução na taxa de acerto das árvores de decisão, entretanto, a redução no índice Kappa foi mais acentuada. Este resultado deve-se a classificação ponderada de cada classe, sendo que a classe de pixels de "café" obteve uma taxa de acerto especifica inferior a classe de "não-café" quando considerado todo o conjunto de teste.

Notou-se uma diminuição na taxa de acerto na árvore com 1500 registros por folha quando comparado a árvore de 7500, e o mesmo para a árvore de 250 registros por folha com a de 750. Pequenas variações na taxa de acerto podem ocorrer e são comuns nestes casos. Isso ocorre pelo fato de que uma árvore maior, ao ramificar um nó, pode gerar uma regra que tenha uma taxa de acerto muito inferior a regra do nó em questão, devido a mudanças sutis nos valores de split dos nós. Caso isso se repita por algumas vezes, pode-se ocorrer da taxa de acerto flutuar em torno de uma faixa para pequenas variações no parâmetro que esta sendo calibrado. Neste caso, notou-se que a taxa de acerto está se estabilizando próxima a faixa de 76% e 77%.

As árvores de decisão não obtiveram resultados melhores do que os trabalhos obtidos em literatura, mesmo comparando a árvore com 250 registros por folha. Seu índice Kappa, de 0,42 foi compatível com o obtido por Marques (2003), mas bastante inferior ao 0,55 obtido por Andrade et al. (2013) para regiões montanhosas.

Tabela 7: Random Forest avaliadas no conjunto de teste.

	Random Forest variando-se o número de árvores na floresta			
	5	10	25	100
Taxa de acerto	81,70%	81,00%	83,97%	84,13%
Índice Kappa	0,54	0,54	0,60	0,60

Novamente, ao se um s modelo no conjunto de treinamento pode-se notar uma redução na taxa de acerto. As Random Forest também apresentaram uma redução mais acentuada no índice Kappa, que previamente chegou a 1,00, mostrando que o modelo havia acertado, praticamente, todos os pixels classificados. Nesta avaliação o índice Kappa ficou muito próximo a 0,6 e foi arredondado para facilitar as comparações. A floresta gerada com 25 árvores não mostrou diferença na classificação, apenas no custo computacional. Ela foi gerada em um tempo cerca de 76% inferior ao tempo utilizado para gerar a floresta com 100 árvores.

Similar ao caso de árvores de decisão tende-se a uma estabilização da taxa de acerto em 84%. Os parâmetros de configuração interna do algoritmo, com exceção do número de árvores na floresta, não se demonstrou eficaz na melhora da taxa de acerto dos modelos.

As Random Forest obtiveram desempenhos mais satisfatórios do que as árvores de decisão, tanto em termos de taxa de acerto quanto índice Kappa. Quando o valor de índice Kappa de 0,60 é comparado à literatura ele pode ser considerado um avanço na área. Andrade et al. (2013) também atingiram este valor de índice, entretanto, foi necessário realizar uma separação de áreas com relevo acentuado e plano. Como este estudo considerou todo o município de Machado (extensão territorial maior) e não fez distinção do relevo das áreas, as random forest podem ser consideradas promissoras para realização de classificação automática de áreas cafeeiras.

5.3 Avaliação de classificadores com seleção de atributos

Na seção anterior as Random Forest mostraram-se melhores classificadores do que as árvores de decisão. A partir desta informação os métodos de seleção de atributos foram aplicados a tal técnica a fim de verificar se haveria um aumento na taxa de acerto dos

classificadores. Foram avaliados quatro métodos de seleção de atributos, o Wrapper, o CFS (*Correlation Feature Selection*), o Infogain e o Gainratio.

O método do Wrapper leva em consideração o algoritmo que estará sendo usado sobre o conjunto de dados, portanto, no caso de uma nova técnica ser utilizada eles devem ser rodados novamente. Wrappers funcionam como uma caixa preta, calculando uma pontuação para um determinado subconjunto. Destes subconjuntos, o mais bem pontuado (que leva a maior taxa de acerto do classificador) é o selecionado. Ao ser aplicado no conjunto de treinamento, o Wrapper selecionou apenas um atributo, que foi a banda 8 (pancromática). Foi gerado um classificador com taxa de acerto de apenas 68%, e os resultados no conjunto de teste foram ainda piores, sendo que sua taxa de acerto caiu para 59,8% com índice Kappa de 0,22. Este caso mostrou que a utilização apenas da banda pancromática não é adequada para estudos relacionados nesta área.

O método de seleção de atributos do CFS seleciona um conjunto aleatório e passa a buscar novos conjuntos com medidas superiores. Este método selecionou quatro atributos para classificação, as bandas 2, 3 5 e 8. O desempenho no conjunto de treinamento foi excelente, obtendo valores similares ao modelo sem seleção gerado com 100 árvores na floresta (taxa de acerto de 99,99% e índice Kappa 1,0). Todavia, ao ser avaliado ao conjunto de teste, o mesmo não manteve seu potencial de predição. Sua taxa de acerto caiu para 78,3% e seu índice Kappa para 0,49. De qualquer maneira, houve uma redução de 20% no tempo de geração deste modelo quando comparado ao modelo gerado com todos os atributos. Caso a demanda computacional de um processo de classificação seja muito grande, pode-se abrir mão de uma taxa de acerto mais alta para que modelos possam ser gerados mais rapidamente.

Além destes, os métodos chamados de Infogain e Gainratio, foram aplicados. Ao contrário dos demais métodos estes realizam um ranqueamento dos atributos com relação a classe. A ordem do ranqueamento dos atributos para estes métodos pode ser encontrada na Tabela 8. A partir deste ordenamento, são testados conjuntos elaborando "linhas de corte", por exemplo, deseja-se cortar os atributos com ganho de informação inferior a $2,0 \times 10^{-2}$, assim seriam selecionadas as bandas 8, 2, 3, 1 e 4.

Tabela 8: Atributos selecionados pelos métodos do InfoGain e Gainratio.

Método de seleção de atributos			
InfoGain	Medida de informação	GainRatio	Ganho de informação
Banda 8	0,14173	Banda 8	$3{,}7335 \times 10^{-2}$
Banda 3	0,11141	Banda 2	$3{,}0976 \times 10^{-2}$
Banda 2	0,10787	Banda 3	$2{,}9092 \times 10^{-2}$
Banda 1	0,09689	Banda 1	$2{,}5554 \times 10^{-2}$
Banda 4	0,09251	Banda 4	$2{,}4206 \times 10^{-2}$
Banda 5	0,04917	Banda 5	$1{,}5524 \times 10^{-2}$
Banda 11	0,04723	Banda 6	$1{,}4316 \times 10^{-2}$
Banda 6	0,04679	Banda 11	$1{,}3702 \times 10^{-2}$
Banda 7	0,04534	Banda 7	$1{,}3534 \times 10^{-2}$
Banda 10	0,04464	Banda 10	$1{,}3361 \times 10^{-2}$
Banda 9	0,00189	Banda 9	$0{,}0953 \times 10^{-2}$

Uma primeira indução foi realizada removendo-se apenas a banda 9, em seguida as bandas 10 e 7, as bandas 11, 6 e 5. Não foi necessário avaliar apenas a banda 8 pois este já foi o conjunto selecionado pelo método do Wrapper, o qual não apresentou bons resultados.

Para a remoção da banda 9, quanto das bandas 9, 10 e 7, os resultados foram praticamente os mesmos, mantendo a taxa de acerto próxima a 84,0% e índice Kappa de 0,60. Quando as bandas 11, 6 e 5 foram removidas houve uma queda acentuada destes valores, que passaram para 81,0% e 0,54. Apesar da baixa na taxa de acerto ela ainda pode ser comparada com o valor obtido por Andrade et al. (2013) para áreas mais acentuadas, que foi de 0,55. Assim pode-se considerar que a exclusão das bandas 7, 9 e 10 não afetou a classificação por Random Forest para áreas cafeeiras.

Ainda foram testadas composições manuais de bandas, como a seleção das bandas azul, verde e vermelho (2, 3 e 4). Destas com adição da banda do infravermelho próximo (banda 5) e da banda pancromática (banda 8), além de uma seleção contendo estas 5 bandas. Todavia, nenhum resultado obtido superou os valores anteriores, sendo assim pode-se concluir que a remoção das bandas 11, 6 e 5 influenciou nos resultados da classificação.

A utilização de poucos atributos no conjunto pode prejudicar a taxa de acerto do algoritmo de Random Forest, fato que ocorre pela dinâmica de funcionamento do mesmo. Um dos primeiros passos deste algoritmo é selecionar atributos aleatoriamente para gerar cada árvore que irá compor a florestas. Quando se há uma diminuição do número de atributos presentes no conjunto de dados, há uma heterogeneidade menor de cada árvore gerada, o que tende a causar uma redução na eficiência das Florestas (WITTEN et al., 2011).

5.4 Geração do mapa temático com áreas classificadas automaticamente

Conforme discutido na seção anterior, o melhor modelo de classificação obteve uma taxa de acerto de 84,13% e índice Kappa de 0,60. A partir dele será gerado um mapa de classificação automática, que poderá ser comparado com a classificação manual realizada. Este é considerado o produto final deste trabalho e pode ser considerado como a última fase do processo KDD, a distribuição. Este mapa será comparado com o mapa de classificação manual gerado por Moreira e adaptado conforme as Figura 15 e 16.

A figura 17 mostra o mapa temático em classificação manual com o mapa de classificação sobrepostos. Este procedimento permitiu a identificação de áreas cafeeiras que foram classificadas incorretamente. Pode-se observar dois pontos, um ao norte e um a sudoeste que apresentaram uma grande quantidade de áreas cafeeiras classificadas incorretamente. Nestes locais do estado há algumas cadeias de morros e, consequentemente, uma declividade que influenciou na precisão da classificação.

Também houve dois pontos a sudeste do estado que obtiveram alta taxa de classificação errada. Estes locais apresentaram outras culturas agrícolas e áreas de mata, as quais foram confundidas com áreas cafeeiras, o que ocasionou a confusão na classificação.

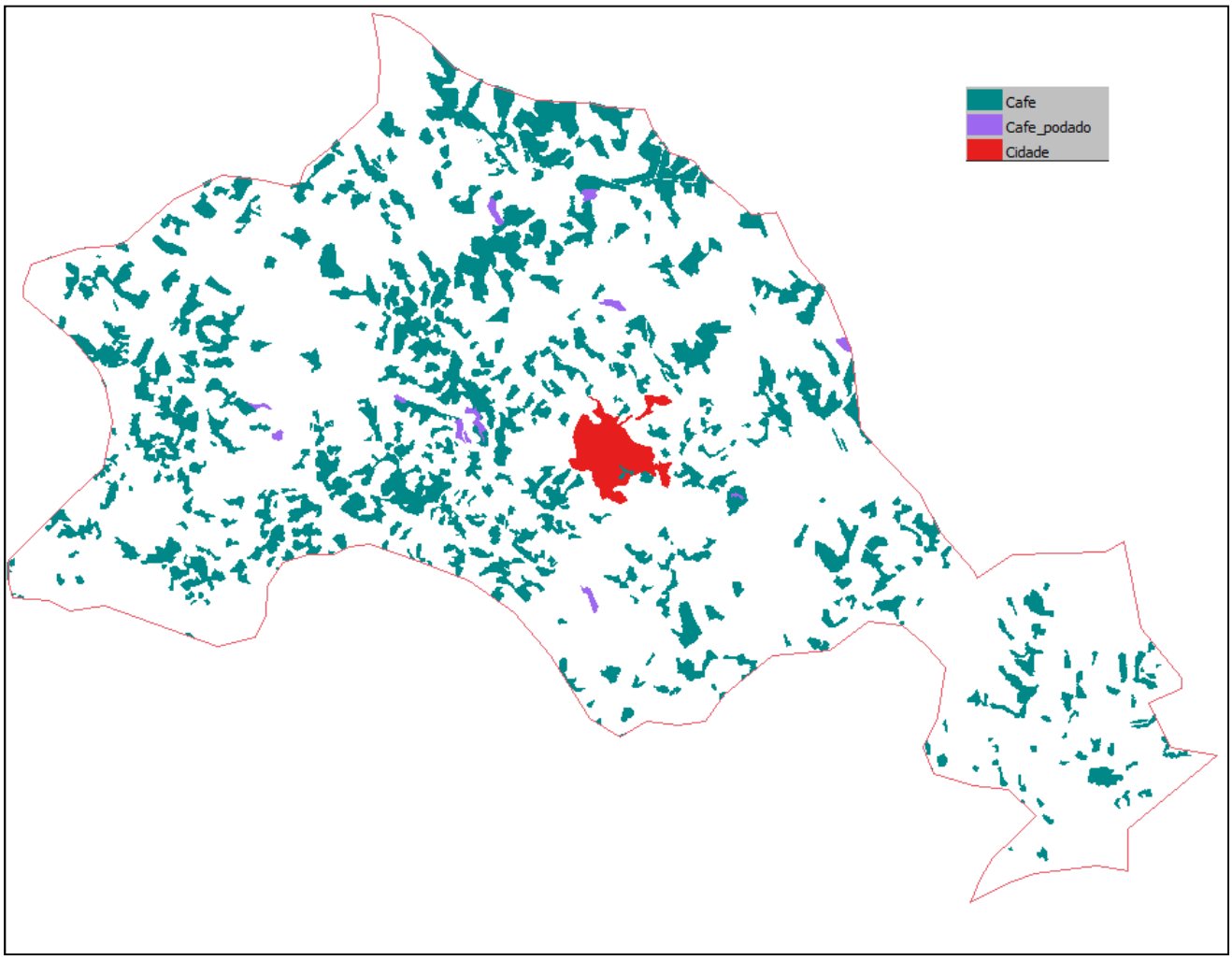

Figura 15: Mapa com a classificação manual de café.

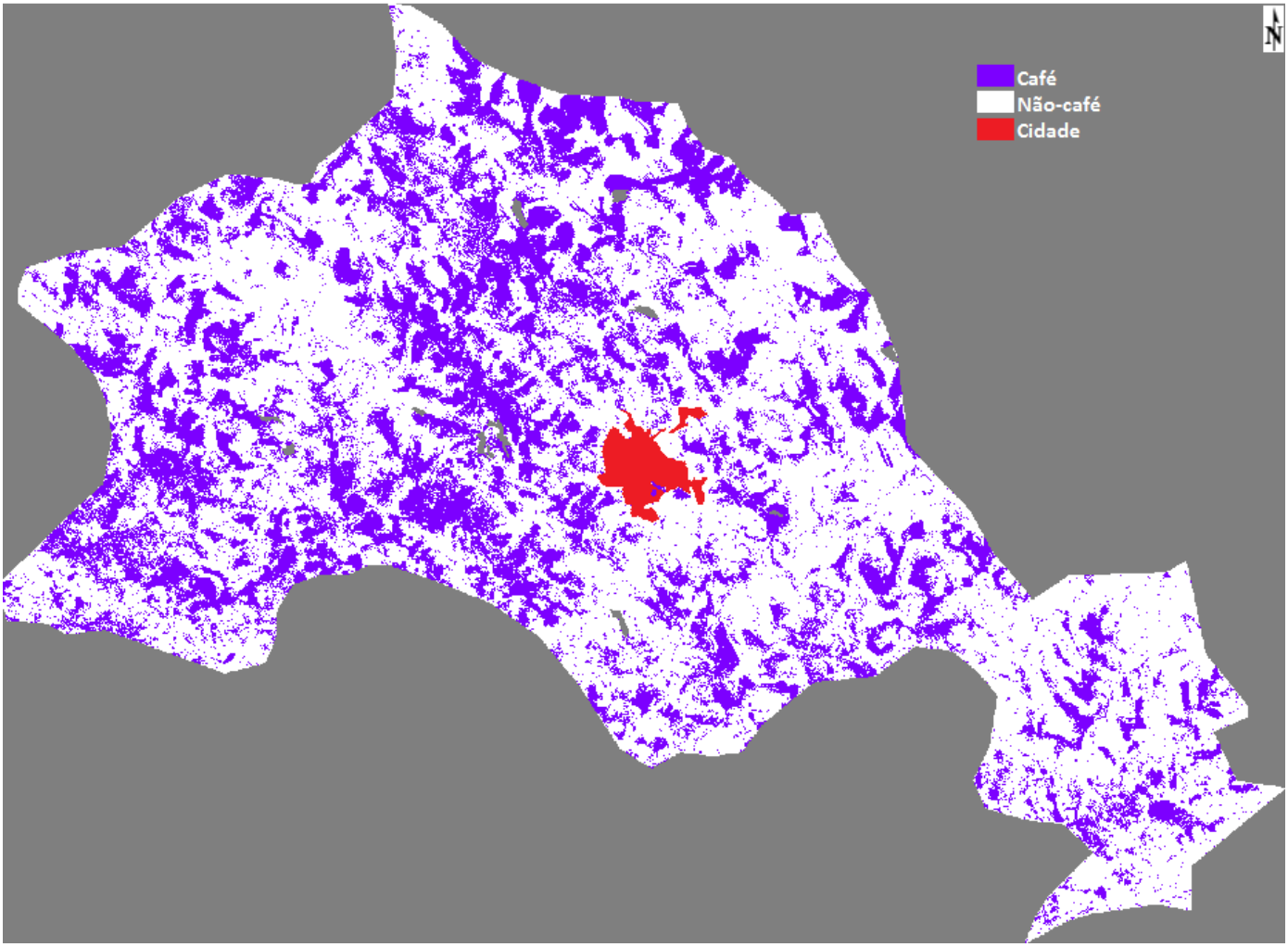

Figura 16: Mapa com a classificação automática de café.

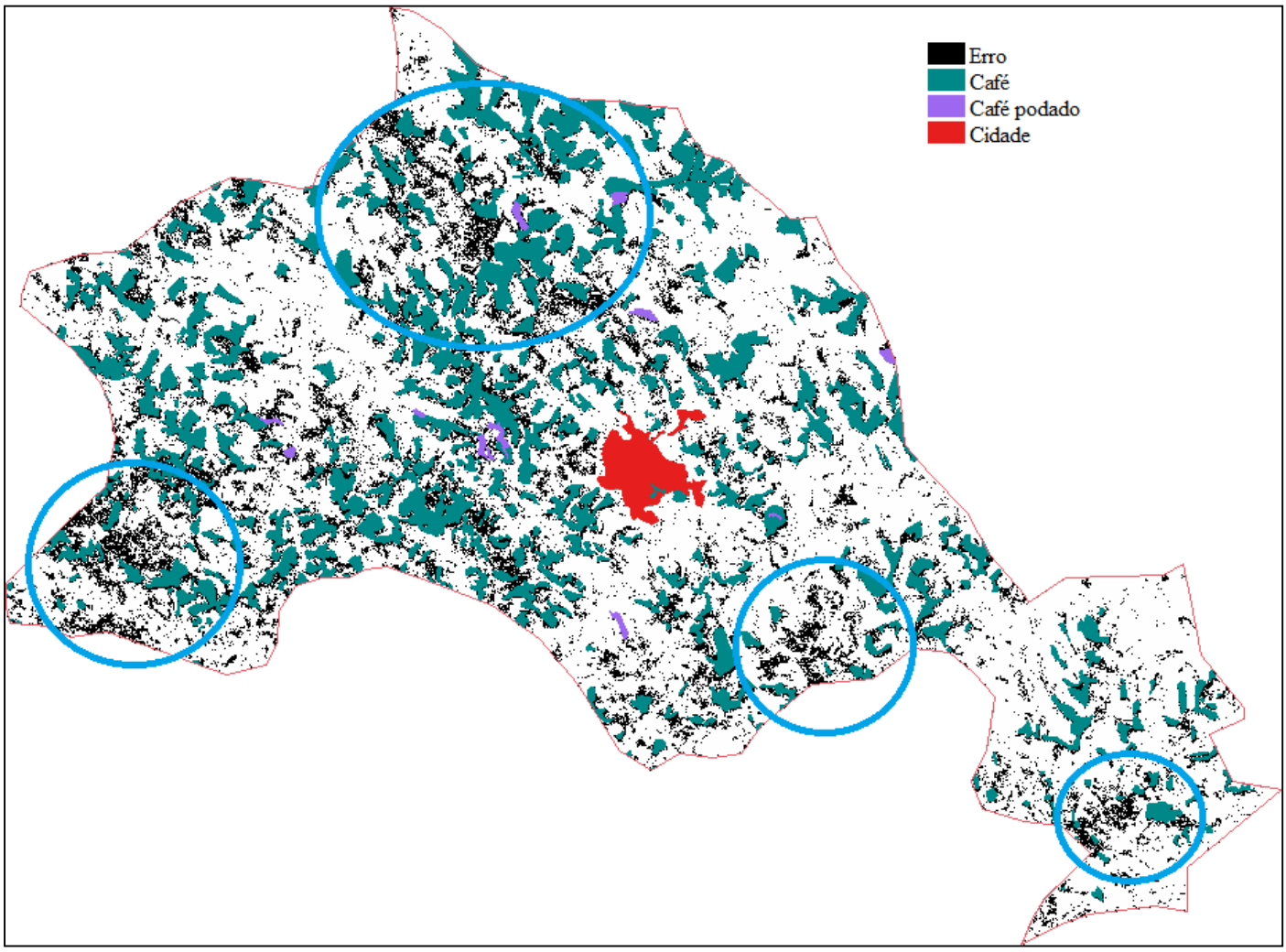

Figura 17: Mapa com a classificação automática de café e os erros na classificação de áreas cafeeiras.

6 Conclusões

A maioria dos classificadores automáticos de imagens espectrais avalia apenas a resposta espectral dos alvos. Quando o alvo estudado, como a cultura cafeeira, apresenta tamanha complexidade, onde fatores ambientais, fenológicos e de manejo interferem na resposta espectral, tais classificadores não apresentam resultados satisfatórios. Assim, Random Forest podem ser consideradas como uma alternativa aos classificadores atualmente utilizados para a classificação de imagens espectrais, por permitirem a utilização de outros parâmetros, além do espectral. Uma grande vantagem destes classificadores automáticos é a rapidez de classificação quando comparado a métodos manuais por exemplo.

Trabalhos futuros visam incorporar à classificação com algoritmos de Random Forest, modificando-se os atributos no conjunto de dados, como a exclusão de bandas não úteis para a classificação e melhor mapeamento da forma e textura dos alvos de treinamento, visto que estes são os atributos utilizados pela interpretação visual. Outra possível alternativa para a melhoria da classificação por Random Forest seria a inserção de outros parâmetros de entrada para o treinamento da Random Forest, como dados de relevo (altitude, orientação de vertente e declive). No Sul de Minas Gerais, onde o café é cultivado em áreas de altitudes mais elevadas, acima de 700m, a altitude pode ser um parâmetro discriminador a ser incorporado à rede, eliminando a possibilidade da Random Forest se confundir.

Os resultados alcançados na região cafeeira de Machado superaram aqueles encontrados na literatura até então, com índice Kappa de 0,60, sendo que os melhores foram de 0,55 e 0,60, obtidos por Andrade et al. (2013). Apesar de ainda não ter sido incorporado à floresta nenhum outro parâmetro, a não ser o espectral, a metodologia proposta separou os ambientes geomorfológicos, a fim de permitir uma melhor classificação; e incluiu máscaras de área urbana e exclusão de áreas de café podado, diminuindo assim a quantidade e a variabilidade dos alvos observados nas imagens.

6.1 Sugestões de trabalhos futuros

Sugerem-se como propostas de continuidade do trabalho os seguintes tópicos:

- Ampliar o número de classes a fim de treinar um modelo com menor variabilidade espectral dos alvos em relação a uma determinada classe.

- Utilizar outros atributos para auxiliar a classificação, tais como textura, forma, relevo, entre outros.

- Testar a metodologia utilizada neste trabalho em imagens com resoluções espaciais melhores;

- Testar a metodologia em imagens com resolução temporal melhor, devido ao ciclo do café;

- Testar a metodologia em outras áreas cafeeiras do Sul de Minas Gerais;

7 Referências bibliográficas

ADAMI, M.; MOREIRA, M. A.; BARROS, M. A.; MARTINS, V. A. Avaliação da exatidão do mapeamento da cultura do café no Estado de Minas Gerais: Atores e Causas da Modificação do Uso do Solo In: Simpósio Brasileiro de Sensoriamento Remoto, 14, 2009, p.1-8, Natal, RN. **Anais...**, 2009.

AGRAWAL, R.; MANILLA, H.; SRIKANT, R.; TOIVONEN, H.; VERKAMO, A. I. Fast Discovery of Association Rules. In: FAYYAD, U.M. PIATETSKY-SHAPIRO, G.; SMYTH, P.; VERKAMO, A.I. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1996, p.307-328.

ANDRADE, L.N.; VIEIRA, T.G.C.; LACERDA, W.S.; DAVIS JUNIOR, C.A. Redes Neurais Artificiais (RNA) aplicadas à classificação de áreas cafeeiras na região de Três Pontas-MG. In: Simpósio Brasileiro de Sensoriamento Remoto, 15, 2011, p.7603-7610, Curitiba, PR. **Anais...**, 2011.

ANDRADE, L.N.; VIEIRA, T.G.C.; LACERDA, W.S.; VOLPATO, M.M.L.; DAVIS JUNIOR, C.A. Aplicação de redes neurais artificiais na classificação de áreas cafeeiras da região de Machado (MG). **Coffee Science,** v.8, n.1, p.78-90, jan./mar. 2013.

APTE, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**. Amsterdam, v. 13, n.2-3, p.197-210, nov., 1997.

BATISTA, G. T.; TARDIN, A. T.; CHEN, S. C. & DALLEMAND, J. F. 1990. Avaliação de produtos HRV/SPOT e TM/Landsat na discriminação de culturas**. Pesquisa Agropecuária Brasileira**, volume 25, n.3, p. 379-386.

BELLE, V. **Detection and recognition of human faces using random forest for a mobile root.** 104p. Dissertação (Mestrado em Ciência da Computação) – Rwthaachen University, Alemanha. 2008.

BERNARDES, T.; ALVES, H. M. R.; VIEIRA, T. G. C. & ANDRADE, H. 2007. Avaliação da acurácia do mapeamento do uso da terra no complexo Serra Negra, Patrocínio, MG. In Simpósio Brasileiro de Sensoriamento Remoto (SBSR), **anais...**, volume 13, Florianópolis. São José dos Campos: INPE. p. 5587-5594.

BISPO, R.C.; LAMPARELLI, R.A.C.; ROCHA, J.V. Using fraction images derived from modis data for coffee crop mapping. **Engenharia Agrícola**, v.34, n.1, p.102–111, 2014.

BRANNSTROM, C.; JEPSON, W.; FILLIPI, A. M.; REDO, D.; XU, S.; GANESH, S. Land change in the Brazilian Savanna (Cerrado), 1986-2002: Comparative analysis and implications for land-use policy. **Land use policy,** v.25, p.579-595, 2008.

BREIMAN, L. Random forests. **Machine Learning Journal**. Hingham, v.45, p.5–32, jan. 2001.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, Charles. J. **Classification and regression trees**. Boca Raton: CRC, 1984.

BURROUGH, P. A. & MCDONNELL, R. A. 1998. **Principles of Geographical Information Systems**. 2 Ed. Oxford, cGraw-Hill Book Company, Cambridge, Massachusetts.

CAMARA, G.; SOUZA, R. C. M.; FREITAS, U. M.; GARRIDO, J. SPRING: Integrating remote sensing and GIS by object-oriented data modelling. **Computers & Graphics**. v. 20, n.3, p. 395-403, Mai/Jun, 1996.

CANDIDO, M. Z.; CALIJURI, M. L.; MOREIRA NETO, R. F. Modelagem do Uso Ocupação e Desenvolvimento de uma região com a ferramenta Land Change Modeler (LCM). In: Congresso Brasileiro de Cartografia, 24, 2010, p.663-668, Aracajú,SE **Anais...** 2010.

CARUANA, R.; KARAMPATZIAKIS, N.; YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. In: International conference on Machine learning, 25, Helsinki, **Proceedings…** Helsinki: ACM, p.96-103, 2008.
CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. Illinois: SPSS, 2000.
COHEN, J**.** A coefficient of agreement for nominal scales. **Eductional and Psychological Measurement**, v.20, n.1, p.37 –46, abr., 1960.
CROS, J.; COMBES, M. C.; TROUSLOT, P.; ANTHONY, F.; HAMON, S.; CHARRIER, A.; LASHERMES, P. Phylogenetic analysis of chloroplast DNA Variation in *Coffea* L. **Molecular Phylogenetics and Evolution** v.9, n.1, p.109– 117, fev., 1998.
DALLEMAND, J. F. 1987. Identificação de culturas de inverno por interpretação visual de dados SPOT e Landsat/TM no Noroeste do Paraná. 131p. **Dissertação** (Mestrado em Sensoriamento Remoto), Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
EMBRAPA. **LANDSAT** - Land Remote Sensing Satellite. Disponível em: <http://www.sat.cnpm.embrapa.br/conteudo/missao_landsat.ph>. Data de acesso: 15/04/2014.
FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v.17, n.3, p.37-54, jul., 1996.
FAZUOLI, L. C.; MEDINA FILHO, H. P.; GONÇALVES, W.; GUERREIRO FILHO, O.; SILVAROLLA, M. B.. Melhoramento do cafeeiro: variedades tipo arábica obtidas no Instituto Agronômico de Campinas. In: ZAMBOLIN, L. **O estado da arte de tecnologias na produção de café.** Viçosa: UFV, 2002. p.163-215.
FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI Magazine**, v.13, n.3, p.57-70, jul., 1992.
GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research.** v.3, p.1157-1182, mar., 2003.
HALL, M. A. **Correlation-based feature selection for machine learning.** 178p. Tese (Doutorado em Ciência da Computação) – Departament of Computer Science, University of Waikato, Nova Zelândia. 1999.
HALL, M. A.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; **SIGKDD Explorations**. New York, v.11, n.1, p. 10-18, jun., 2009.
HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.
HELFAND, S. M.; REZENDE, G. C. Padrões Regionais de Crescimento da produção de grãos no Brasil e Papel da Região Centro-Oeste. Rio de Janeiro: IPEA, 2000
INPE. **Instituto Nacional de Pesquisas Espaciais**.
< http://www.dgi.inpe.br/Suporte/files/Cameras-LANDSAT57_PT.php >, 20/03/2013.
JENSEN, J. R. **Remote Sensing of the environment**: an earth resource perspective. 2ed. San Francisco: Prentice Hall, 2011
JOHN, G. H.; KOHAVI, R. Wrappers for feature subset selection. **Artificial Inteligence**. v.97, n.1-2, p.273-324, dez., 1997.
LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, n.1, p.159-174, mar., 1977.
LILLESANT, T.M. **Remote sensing and image interpretation**, 2. ed. New York: John Wiley, 1987. 721 p.
MACHADO, M. L. 2002. Caracterização de agroecossistemas cafeeiros da Zona da Mata de Minas Gerais, usando Sensoriamento Remoto e Sistemas de Informações Geográficas, 137p.

Dissertação (mestrado em solos e nutrição de plantas), Universidade Federal de Lavras, Lavras.
MARQUES, H. S. 2003. Uso de geotecnologias no estudo das relações entre solos, orientação de vertentes e o comportamento espectral de áreas cafeeiras em Machado, Minas Gerais. 82p. **Dissertação** (mestrado em solos e nutrição de plantas), Universidade Federal de Lavras, Lavras.
MATIELLO, J. B.; SANTINATO, R.; GARCIA, A. W. R.; ALMEIDA, S. R.; FERNANDES, D. R. Cultura de café no Brasil. Novo manual de recomendações. In: MATIELLO, J. B. **MAPA/PROCAFÉ**. Rio de Janeiro, p.387, 2002.
MENDES, L. C.; MENEZES, H. C.; SILVA, M. A. A. P. Optimization of the roasting of robusta coffee (*C. canephora* conillon) using acceptability testes and RSM. **Food Quality and Preference.** v.12, n.2, p.153-162, 2001.
MINISTÉRIO DA AGRICULTURA. **Ministério da Agricultura, Pecuária e abastecimento.** <www.agricultura.gov.br.>, 26/01/2013.
MINISTÉRIO DO DESENVOLVIMENTO. **Ministério do desenvolvimento, indústria e comércio exterior.** <http://www.mdic.gov.br/sitio/interna/interna.php?area=5&menu=1955>, 20/03/2013.
MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 89-135.
MOREIRA, M. A. 2001. **Fundamentos do sensoriamento remoto e metodologias de aplicação**. São José dos Campos: INPE. 249p.
MOREIRA, M. A. 2003. **Fundamentos do sensoriamento remoto e metodologias de aplicação**. 2 Ed. Viçosa: UFV. 307p.
MOREIRA, M. A.; ADAMI, M.; RUDORFF, B. F. T. Análise espectral e temporal da cultura do café em imagens Landsat. **Pesquisa Agropecuária Brasileira**, v.39, n.3, p.223-231, mar. 2004.
MOREIRA, M. A.; BARROS, M. A.; ROSA, V. G. C.; ADAMI, M. Tecnologia de informação: imagens de satélite para o mapeamento de áreas de café de Minas Gerais. **Informe Agropecuário**, v.28, n.241, 2007.
MOREIRA, M. A.; BARROS, M. A.; RUDORFF, B. F. T. Geotecnologias no mapeamento da cultura do café em escala municipal. **Sociedade & Natureza**, v.20, n.1, p.101-110, jun. 2008
NOVO, E. M. L. M. 1989. **Sensoriamento Remoto: Princípios e Aplicações**. São Paulo: Edgard Blücher. 308p.
PONZONI, F. J.; SHIMABUKURO, Y. E. **Spectral properties of vegetation**. Material do curso de comportamente espectral de alvos. INPE, São José dos Campos, 1991, 15 p.
PROCAFÉ. Fundação PROCAÉ Minas Gerais.
< http://www.fundacaoprocafe.com.br/sites/default/files/publicacoes/pdf/revista/Coffea11.pdf >, 20/03/2013.
PRUDENTE, T. D.; ROSA, R. Geoprocessamento e sensoriamento remoto aplicados no mapeamento do uso da terra e cobertura vegetal no município de Tucaciguara-MG In: Simpósio Brasileiro de Geografia Física Aplicada, 12, 2007, Natal, RN. **Anais...**, 2007.
RAMIREZ, G. M.; ZULLO JUNIOR, J.; ASSAD, E. D.; PINTO, H. S. Comparação de dados dos satélites Ikonos-II e Landsat/ETM+ no estudo de áreas cafeeiras. **Pesquisa Agropecuária Brasileira**, v.41, n.4, 2006.
REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; DE PAULA, M. F. Mineração de Dados. In: REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri:

Editora Manole, 2002. p. 89-135.
ROSA, R. **Introdução ao sensoriamento remoto.** 7.ed. Uberlândia: EDUFU, 2009.
SANTOS, A. B.; PETRONZIO, J. A. C. Mapeamento de uso e ocupação do solo do município de Uberlândia-MG utilizando técnicas de Geoprocessamento. In: Simpósio Brasileiro de Sensoriamento Remoto, 15, 2011, p.6185-6192, Curitiba, PR. **Anais...**, 2011.
USDA. **United States Department Of Agriculture.** <www.usda.gov>, 15/02/2013.
VIEIRA, T. G. C.; ALVES, H. M. R.; LACERDA, M. P. C.; VEIGA, R. D. & EPIPHANIO, J. C. N. 2006. Crop parameters and spectral response of coffee (coffea arábica l.) areas within the state of Minas Gerais, Brazil. **Coffee Science**, volume 1, n.2, p. 111-118.
WEISS, G. M.; PROVOST, F. **The effect of class distribution on classifier learnig**: an empirical study. Techinical Report ML-TR-44, Departamente of computer science, Rutgers University, 2001.
WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3ed. San Francisco: Morgan Kaufmann, 2011.