BCG - Boletim de Ciências Geodésicas - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702016000400043

Artigo

ALGORITMOS DE APRENDIZAGEM DE MÁQUINA E VARIÁVEIS DE SENSORIAMENTO REMOTO PARA O MAPEAMENTO DA CAFEICULTURA

Machine learning algorithms and variable of remote sensing for coffee cropping mapping

Carolina Gusmão Souza [1]

Luis Carvalho [1,*]

Polyanne Aguiar [2]

Tássia Borges Arantes [1]

[1]Departamento de Ciências Florestais, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil; carolinagusmaosouza@gmail.com; tassiabarantes@gmail.com.

[2]Departamento de Ecologia, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil; polyanneaguiar@gmail.com.

*****Autor de correspondência: Luis Carvalho, passarinho@dcf.ufla.br

Resumo:

A cafeicultura é uma das principais culturas agrícolas do Brasil e realizar o mapeamento e monitoramento desta cultura é fundamental para conhecer sua distribuição espacial. Porém, mapear estas áreas utilizando imagens de Sensoriamento Remoto não é uma tarefa fácil. Sendo assim, este trabalho foi realizado com o objetivo de comparar o uso de diferentes variáveis e algoritmos de classificação para o mapeamento de áreas cafeeiras. O trabalho foi desenvolvido em três áreas diferentes, que são bastante significativas na produção de café. Foram utilizados 5 algoritmos de aprendizagem de máquinas e 7 combinações de variáveis: espectrais, texturais e geométricas, associadas ao processo de classificação. Um total de 105 classificações foram realizadas, 35 classificações para cada uma das áreas. As classificações que não usaram variáveis espectrais não resultaram em bons índices de acurácia. Nas três áreas, o algoritmo que apresentou as melhores acurácias foi o S*upport vector machine*, com acurácia global de 85,33% em Araguari, 87% em Carmo de Minas e 88,33% em Três Pontas. Os piores resultados foram encontrados com o algoritmo *Random Forest* em Araguari, com acurácia global de 76,66% e com o *Naive Bayes* em Carmo de Minas e Três Pontas, com 76% e 82% de acerto. Nas três áreas, variáveis texturais, quando associadas às espectrais, melhoraram a acurácia da classificação. O SVM apresentou o melhor desempenho para as três áreas.

***Palavras-chave*:** classificação automatizada; sensoriamento remoto; algoritmos de aprendizagem de máquina, cultura cafeeira.

Abstract

Coffee is one of the main crops in Brazil, therefore, performing the mapping and monitoring of this culture is essential for know your special distribution. However, map this culture is not an

easy task. Thus, the objective of this study was to compare the use of different variables and classification algorithms for coffee area classification. The study was conducted in three areas, environmentally different. We use 5 machine learning algorithms and 7 combinations of variables, using spectral, textural and geometric variables associated with the classification process. A total of 105 maps were made. All ratings that have not used spectral variables don't achieved good levels of accuracy. In all three areas, the algorithm that presented the best accuracies was the Support Vector Machine with overall accuracy 85.33% in Araguari, 87.00% in Carmo de Minas and 88.33% in Três Pontas. The worst results were found by Random Forest algorithm in Araguari, with 76.66% accuracy and Naive Bayes in Carmo de Minas and Três Pontas, with 76.00% and 82.00%. In all three areas, textural variables when associated with spectral, improved the classification accuracy. The SVM showed the best performance for the three areas.

Keywords: automatized mapping; remote sensing; machine learning algorithms, coffee cropping.

1. Introdução

A cafeicultura é uma importante atividade econômica no panorama internacional (OIC, 2014). A maior movimentação ocorre nos Estados Unidos, que comercializa cerca de 23 milhões de sacas/ano (OIC, 2014). O café é uma das principais culturas agrícolas do Brasil, com grande importância para a economia do país (CONAB, 2014). Minas Gerais se destaca como o maior estado produtor de café, com mais de 50,00% da produção nacional (CONAB, 2014). Projeções indicam um crescimento de 2,00% ao ano, em relação ao total de café que é consumido atualmente (OIC, 2014). Sendo assim, torna-se fundamental conhecer a distribuição espacial da atividade cafeeira para prever e planejar seu crescimento, bem como a estratégia de comercialização da sua produção de forma eficiente.

Tecnologias e sistemas associados ao Sensoriamento Remoto têm sido amplamente empregadas para mapear e monitorar áreas agrícolas (Veloso, 1974; Moreira et al., 2004; Cordero-Sancho & Sader, 2007; Li et al., 2014). Existem, no entanto, diversas dificuldades que envolvem o processo de derivar informações úteis a partir de imagens de Sensoriamento Remoto. Um dos complicadores é a heterogeneidade de paisagens dos ambientes tropicais (Li et al., 2014) que, consequentemente, aumenta a complexidade das cenas retratadas nas imagens, em termos de elementos registrados. Essa complexidade, por sua vez, aumenta a confusão espectral entre os diferentes tipos de cobertura da terra com respostas espectrais semelhantes, como é o caso das áreas cobertas por cafezais e por remanescentes de vegetação nativa (Moreira et al., 2004; Cordero-Sancho & Sader, 2007; Adami et al., 2009).

Estas limitações afetam diretamente o mapeamento dos cafezais devido à ampla variabilidade espectral, temporal e espacial das lavouras cafeeiras (Vieira et al., 2007; Adami et al., 2009), tornando a definição de um padrão de identificação para o café mais difícil do que para outras culturas agrícolas. Segundo Adami et al. (2009), os diversos métodos de planejamento do plantio, como espaçamento e sistema de cultivo, apresentam similaridade espectral com áreas de mata nativa e com outras culturas. O mapeamento pode ser dificultado, ainda, em regiões montanhosas (Santos et al., 2012; Andrade et al., 2013a). Estas áreas sombreadas aumentam ainda mais a variabilidade espectral da cultura nas imagens de Sensoriamento Remoto (Santos et al., 2012).

O mapeamento de áreas agrícolas usando dados de Sensoriamento Remoto, especialmente aquelas relacionadas à cafeicultura, já vem sendo pesquisado desde a década de 1970 (Veloso, 1974), porém, os resultados deste primeiro estudo não permitiam uma discriminação apropriada dos cafezais em relação a outros tipos de uso. Em Minas Gerais, Moreira et al. (2004), Vieira et al. (2007), Machado et al. (2010), Santos et al. (2012) e Andrade et al. (2013a) trabalharam mais recentemente no mapeamento e na caracterização desta cultura e reportaram resultados promissores, porém, apresentando muita confusão entre os alvos. Andrade et al. (2013a) conduziram uma classificação automática para mapear uma área cafeeira na região de Machado, MG. A área foi dividida em duas, uma com relevo mais suave e a outra com o relevo mais movimentado. Os resultados mostraram índice *Kappa* inferiores a 0,60. Cordero-Sancho & Sader (2007) mapearam áreas de café na Costa Rica e analisaram a combinação de bandas espectrais, além de dados complementares, a fim de avaliar a precisão do mapeamento desta cultura e de outros tipos de cobertura da terra. O maior índice de acurácia global foi de 65,00%, tendo a separação espectral entre floresta e café não sido bem sucedida. Avaliando a importância de variáveis espectrais, geométricas e texturais para o mapeamento de áreas urbanas, Wieland et al. (2014), demonstraram que as variáveis mais importantes no processo de classificação foram as espectrais, seguidas das texturais e, por fim, das geométricas.

Estudos realizados na América Latina utilizaram apenas métodos de classificação mais tradicionais, como a classificação baseada em pixel e classificadores paramétricos padrões, como por exemplo, máxima verossimilhança (Adami et al., 2009; Martínez-Verduzco et al., 2012). Poucos estudos têm utilizado a classificação orientada a objetos combinados com novos algoritmos de classificação (Santos et al., 2012), como os algoritmos de aprendizagem de máquina (AM).

Algoritmos de aprendizagem de máquina, como *Suport vector machine* (SVM), mostraram bons resultados na acurácia do mapeamento do café (Santos et al., 2012; Sarmiento et al., 2014). Porém, poucos estudos trabalharam com este algoritmo para o mapeamento do café no Brasil (Santos et al., 2012; Marujo et al., 2013). Além disso, alguns algoritmos de AM que vêm sendo muito utilizados em classificação de imagens ainda não foram utilizados para o mapeamento da cafeicultura, como é o caso dos algoritmos *Random Forest* (RF), *Decision Tree* (DT) e *Naive Bayes* (NB). Estes algoritmos têm demonstrado excelente desempenho na análise de bases de dados de Sensoriamento Remoto que apresentam grande complexidade (Li et al., 2013).

Alguns estudos também têm incorporado variáveis geométricas e texturais para auxiliar o mapeamento de lavouras cafeeiras, mas ainda não se sabe se estas variáveis são eficientes para distinguir as plantações de café de outros tipos de uso da terra, como vegetação e pastagem (Gomez et al., 2010; Santos et al., 2012; Marujo et al., 2013). Estudos analisando estas variáveis separadamente e em conjunto ainda não foram realizados para o mapeamento desta cultura.

É importante salientar, ainda, que existe uma escassez de trabalhos utilizando imagens de alta resolução espacial para o estudo da cafeicultura (Marujo et al., 2013; Sarmiento et al., 2014) e para os demais mapeamentos do uso e cobertura da terra. Na maioria dos casos são utilizadas imagens de média resolução espacial, como, por exemplo, imagens dos sensores TM e ETM+ (Wieland et al., 2014; Pradhan et al., 2013; Otukei & Blaschke, 2010).

Sendo assim, considerando a importância da produção cafeeira para a economia do país, a dificuldade intrínseca de mapeamento desta cultura e o fato de ainda não existir um mapa oficial dos parques cafeeiros do Brasil, fica evidente a necessidade de mais estudos para subsidiar o desenvolvimento de métodos apropriados e confiáveis de mapeamento das lavouras de café.

O estudo apresentado visa contribuir para o conhecimento das peculiaridades do mapeamento de cafezais, analisando o desempenho de algoritmos inovadores e de variáveis que retratam características espectrais, geométricas e texturais na classificação digital de imagens de alta

resolução espacial. Este trabalho foi realizado com o objetivo principal de comparar algoritmos de AM usando diferentes conjuntos de variáveis derivadas de dados de Sensoriamento Remoto e identificar as melhores combinações algoritmos-variáveis para o mapeamento de cafezais em três regiões de Minas Gerais.

Nesse sentido, pretendeu-se responder às seguintes questões: (1) Há diferença significativa entre as classificações utilizando diferentes combinações entre algoritmos de classificação e variáveis de entrada? (2) Quais conjuntos de variáveis são mais eficientes para classificar cafezais? (3) Quais são os algoritmos mais eficientes para classificar cafezais? Qual algoritmo é mais acurado para o mapeamento de cafezais em cada uma das três regiões em estudo? (4) Quais são as classes confundidas com cafezal durante o processo de classificação?

2. MATERIAIS E MÉTODOS

2.1 Área de estudo e dados de Sensoriamento Remoto

As áreas de estudo estão localizadas no estado de Minas Gerais, Brasil. Foram escolhidas três áreas distintas, denominadas áreas I, II e III: a primeira corresponde à cena RapidEye 2230526, registrada em 14/06/2010, que cobre parte do município de Araguari (área I), região oeste do estado. As outras duas áreas estão localizadas na região sul do estado e correspondem às cenas RapidEye 2328914, registrada em 22/07/2010 e 2329213, registrada em 18/06/2010, cobrindo parte dos municípios de Carmo de Minas (área II) e Três Pontas (área III), respectivamente, conforme Figura 1. Na região sul, o bioma predominante é Mata Atlântica e, na região oeste, é o bioma Cerrado.

O clima no município de Araguari, de acordo com a classificação de Köppen, é do tipo Cwa, tropical de altitude, com temperatura média de 21ºC e índice pluviométrico anual de 1.400 mm. A altitude média é de 1.013 m, onde são encontradas as formas tabulares, e as atividades principais são a agricultura de grãos (soja e milho) e a cafeicultura (IBGE, 2009).

O município de Carmo de Minas tem altitude média de 960 m, temperatura média anual de 19,1ºC e índice pluviométrico médio anual de 1.568 mm (IBGE, 2009). O clima, segundo a classificação de Köppen, é Cwb, subtropical de altitude, com temperatura média anual de 17ºC. A base da sua economia é a agricultura, destacando-se a cafeicultura e a pecuária (IBGE, 2009).

O município de Três Pontas tem altitude média de 905 m. O clima, de acordo com a classificação de Köppen, é tropical de altitude. A temperatura média anual de 18ºC e média anual de pluviosidade é de 1.440 mm. A principal atividade econômica da região é a cafeicultura (IBGE, 2009).

Estas regiões foram escolhidas por serem áreas com fitofisionomias e características ambientais diferentes e também por estarem localizadas em municípios representativos na produção de café no cenário mineiro (CONAB, 2014; Souza et al., 2012).

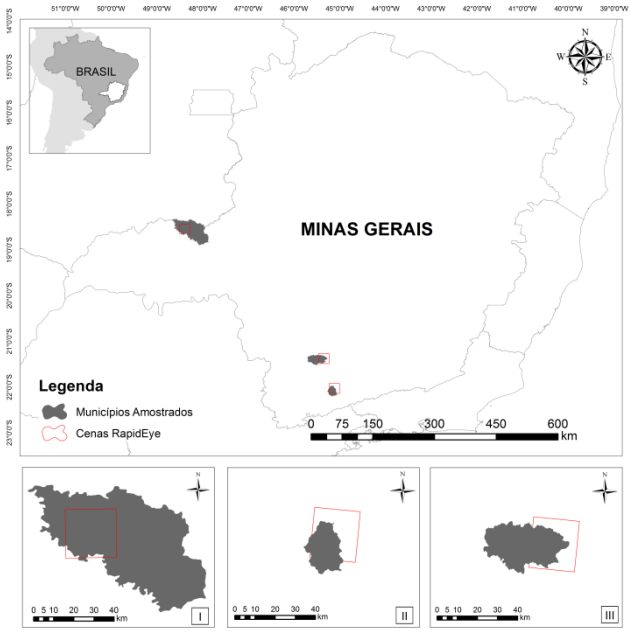

Figura 1: Mapa de localização. Na região oeste do estado está o município de Araguari (área I) e na região sul, os municípios de Carmo de Minas (área II) e Três Pontas (área III).

As imagens selecionadas para este estudo são do satélite RapidEye, as quais têm 5 m de resolução espacial, resolução radiométrica de 16 bits e resolução espectral de cinco bandas. As imagens RapidEye utilizadas neste estudo foram adquiridas com correções geométricas e radiométricas (nível 3A).

2.2 Estratégia de classificação

Todas as imagens RapidEye foram segmentadas utilizando o algoritmo multirresolução, do *software* eCognition Developer, em que foram testados diferentes parâmetros de escala, compacidade e forma, a fim de verificar quais parâmetros de segmentação foram mais adequados, considerando as áreas de café e suas variações na paisagem. Os parâmetros de segmentação escolhidos foram: escala 250 e, para forma e compacidade, foi utilizado o peso 0,5 para ambas. Estas características foram escolhidas por apresentarem, visualmente uma boa separação das áreas de café para as três áreas de estudo, em relação a outros parêmtros testados.

Foram definidas cinco classes para as classificações, de acordo com as suas características espectrais, sendo elas: vegetação nativa: áreas de formações florestais densas e florestas de galeria, e formações de cerrado; Café (subdividida em 3 classes), sendo elas: Café 1: lavouras em idade não produtiva, no início do estágio de crescimento; Café 2: lavouras em estágio de crescimento intermediário; Café 3: lavouras com idade superior a 3 anos; Pastagem: áreas de pastagens naturais e formadas; Outros usos (subdividida em 2 classes), sendo elas: Outros usos 1: áreas com culturas anuais em diversos estágios de desenvolvimento; floresta de produção; Outros usos 2: áreas urbanas e benfeitorias, áreas de solo exposto, áreas de queimadas; e Corpos d'água: rios, córregos e represas, lagoas naturais e artificiais. Como ilustrado na Figura 2.

Para cada área, foram selecionados, no mínimo, 10% do total de objetos gerados pela segmentação para servirem de amostras de treinamento, de acordo com metodologia proposta por Neil et al. (2005). Isso gerou um total de 751 amostras para Araguari, 1.058 para Carmo de

Minas e 939 para Três Pontas. As amostras de treinamento foram coletadas de maneira uniforme por toda imagem. A escolha dos objetos foi feita utilizando-se interpretação visual e de forma criteriosa. Em cada uma das áreas, as amostras escolhidas foram empregadas em todas as classificações.

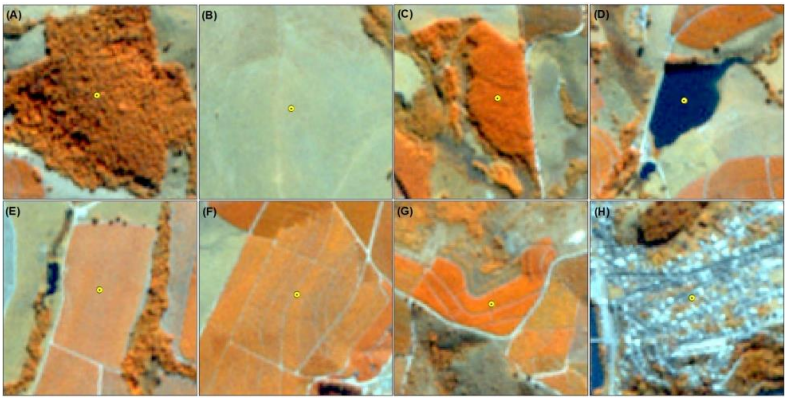

Figura 2: Exemplos de amostras de treinamento. a) vegetação nativa; b) pastagem; c) outros usos 1; d) corpos d'água; e) café 1; f) café 2; g) café 3; h) outros usos 2.

2.3 Processo de classificação

A fim de avaliar o desempenho de diferentes métodos de classificação da cobertura da terra, foram utilizados os seguintes algoritmos: *decision tree* (DT), *naive bayes* (NB), *randon forest* (RF), *support vector machine* (SVM) e *K-nearest neighbor* (KNN).

Estes algoritmos foram utilizados por estarem disponíveis para qualquer usuário e serem de fácil utilização. Todas as fontes de códigos são pacotes oriundos do software RStudio. Os parâmetros empregados pelos algoritmos de classificação estão descriminados na Tabela 1. Para escolher os parâmetros a serem usados no treinamento dos algoritmos, foram realizados testes preliminares e escolhidas as combinações de parâmetros que apresentaram maior precisão durante os ajustes dos modelos.

Tabela 1: Conjunto de parâmetros utilizados em cada algoritmo e sua fonte de códigos.

Algoritmo	Sigla	Fonte de códigos	Tipos de parâmetros	Descrição dos parâmetros	Parâmetros
Decision tree	DT	rpart	cp	É usado para controlar o tamanho da árvore de decisão e para selecionar o tamanho ideal da árvore.	0.001
			maxdepth	Define a profundidade máxima de qualquer nó da árvore final.	30
K-*nearest neighbor*	KNN	eCognition	k	São identificados por uma medida de distância que compara os vetores de características da instância não marcada e o conjunto de instâncias de treinamento fornecido ao classificador. Depois que uma lista de vizinhos mais próximos é obtida, a previsão é baseada em votação (maioria ou distância ponderada).	1
Naive bayes	NB	eCognition	laplace	Suaviza dados categóricos, com o propósito de evitar que o cálculo da probabilidade seja igual à zero.	0
Random forest	RF	randomForest	mtry	Número de variáveis amostradas aleatoriamente como candidatos em cada divisão. Note-se que os valores padrões são diferentes para classificação.	2, 4, 6, 8, 10, 12, 14, …, 50
			ntree	Número de árvores.	1000
Support vector machine	SVM	Kernlab	σ	O parâmetro σ define a largura da função de kernel.	0.02
			kernel	Define qual função kernel é utilizada no treinamento e na predição dos dados.	rbfdot - Radial Basis Function
			C	Ajusta a sensibilidade da margem de decisão de vetores de suporte erroneamente classificados.	1

Tabela2: Relação das variáveis usadas neste estudo. Variáveis relacionadas às informações espectrais, geométricas e texturais.

Espectrais	Brilho	**Texturais**	GLCM* Correlação banda 1
	Desvio padrão banda 1		GLCM Correlação banda 2
	Desvio padrão banda 2		GLCM Correlação banda 3
	Desvio padrão banda 3		GLCM Correlação banda 4
	Desvio padrão banda 4		GLCM Correlação banda 5
	Desvio padrão banda 5		GLCM Desvio padrão banda 1
	Máxima diferença		GLCM Desvio padrão banda 2
	Média banda 1		GLCM Desvio padrão banda 3
	Média banda 2		GLCM Desvio padrão banda 4
	Média banda 3		GLCM Desvio padrão banda 5
	Média banda 4		GLCM Entropia banda 1
	Média banda 5		GLCM Entropia banda 2
	NDVI		GLCM Entropia banda 3
	SAVI		GLCM Entropia banda 4
Geométricas	Área		GLCM Entropia banda 5
	Assimetria		GLCM Homogeneidade banda 1
	Circularidade		GLCM Homogeneidade banda 2
	Compacidade		GLCM Homogeneidade banda 3
	Comprimento		GLCM Homogeneidade banda 4
	Comprimento / Largura		GLCM Homogeneidade banda 5
	Comprimento da borda		GLCM Média banda 1
	Densidade		GLCM Média banda 2
	Índice de borda		GLCM Média banda 3
	Índice de forma		GLCM Média banda 4
	Largura		GLCM Média banda 5

*GLCM (Grey Level Co-occurrence Matrix)

Alguns algoritmos de AM mostram as variáveis mais importantes utilizadas para o processo de classificação. Para selecionar as melhores variáveis utilizadas, foi utilizado o ranking obtido pelo *random forest* e pelo *decision tree*.

2.3 Amostras de acurácia e análise estatística

Com a finalidade de verificar a acurácia das classificações, foram coletados 300 pontos como dados de validação, para cada uma das áreas. Na área I, os pontos coletados foram divididos da seguinte forma: 60 como pastagem, 90 como vegetação, 70 como outros usos, 65 como café e 15 como água. Na área II: 95 como pastagem, 80 como vegetação, 40 como outros usos, 70 como

café e 15 como água. Na área III: 75 como pastagem, 70 como vegetação, 40 como outros usos, 100 como café e 15 como água. Esses pontos foram obtidos por meio de uma amostragem estratificada aleatória, em que cada estrato foi representado por uma categoria de classificação (Martínez-Verduzco et al., 2012). A conferência foi realizada a partir de visitas a campo e do aplicativo Google Earth. Foram utilizados o índice *Kappa* e a acurácia global como critério de avaliação.

Para comparar os resultados das acurácias foi utilizado o teste de McNemar (Foody, 2004), um teste não paramétrico e que avalia a significância estatística das diferenças entre as duas classificações, baseando-se em matrizes de confusão de duas dimensões. Neste teste, a atenção é focada na distinção binária entre a alocação das classes corretamente ou incorretamente. O teste de McNemar é baseado no teste estatístico normal padronizado (Foody, 2004).

3. RESULTADOS E DISCUSSÃO

Neste trabalho avaliou-se a efetividade do uso de diversos classificadores, bem como de diferentes variáveis para a classificação de áreas cafeeiras, visando melhorar a separabilidade entre as classes.

Foram utilizados 5 algoritmos de AM e 7 combinações de variáveis, portanto, um total de 105 classificações foram gerados para as três áreas (35 para cada uma delas). As classificações que apresentaram índice *Kappa* inferior a 0,65 não foram considerados nas análises subsequentes. As classificações feitas usando variáveis apenas texturais e geométricas, ou seja, sem as variáveis espectrais, apresentaram índice *Kappa* inferiores. Sendo assim, optou-se por retirar estas classificações das análises, uma vez que apresentaram resultados considerados insatisfatórios, segundo a classificação de Landis & Koch (1977), portanto, foram consideradas nos resultados e discussão apenas as classificações que incluíam valores acima de 0,65, mostrado na Tabela 3.

Tabela 3: Valores de índice *Kappa*, acurácia global e porcentagem de acerto para cada classe de uso. Estes valores estão dispostos para cada área, cada algoritmo usado e cada conjunto de variáveis. Área I (Araguari), área II (Carmo de Minas) e área III (Três Pontas).

Área	Algoritmos	Variáveis	Acurácia						
			Global (%)	Kappa	Vegetação (%)	Pastagem (%)	Café (%)	Outros usos (%)	Água (%)
I	DT	stg	81.33	0.75	84.44	91.67	92.31	57.14	86.67
		s	82.66	0.77	86.67	91.67	93.95	58.57	86.67
		sg	81.33	0.75	84.44	91.67	92.31	57.14	86.67
		st	81.33	0.75	84.44	91.67	90.77	58.57	86.67
	KNN	stg	83.00	0.77	91.11	93.33	90.77	52.86	100.00
		s	81.66	0.76	88.89	91.67	93.85	50.00	93.33
		sg	81.66	0.76	91.11	93.33	90.77	48.57	93.33
		st	82.33	0.76	91.11	88.33	89.23	58.57	86.67
	NB	stg	79.66	0.73	93.33	95.00	90.77	44.29	53.33
		s	83.66	0.78	88.89	91.67	98.46	57.14	80.00
		sg	83.00	0.77	92.22	95.00	93.85	52.86	73.33
		st	81.33	0.75	92.22	91.67	95.38	51.43	53.33
	RF	stg	78.33	0.71	78.89	93.33	93.85	50.00	80.00
		s	76.66	0.69	70.00	91.67	95.38	52.86	86.67
		sg	77.33	0.70	78.89	91.67	87.69	52.86	80.00
		st	80.00	0.74	78.89	91.67	95.38	54.29	93.33
	SVM	stg	82.66	0.77	91.11	93.33	92.31	50.00	100.00
		s	85.33	0.80	93.33	93.33	95.38	57.14	93.33
		sg	80.33	0.74	91.11	81.67	95.38	47.14	100.00
		st	84.66	0.80	93.33	91.67	95.38	55.71	93.33
Área	Algoritmos	Variáveis	Acurácia						
			Global (%)	Kappa	Vegetação (%)	Pastagem (%)	Café (%)	Outros usos (%)	Água (%)
II	DT	stg	84.33	0.78	71.25	98.95	91.43	67.50	73.33
		s	84.33	0.78	75.00	98.95	85.71	67.50	80.00
		sg	84.33	0.78	75.00	98.95	85.71	67.50	80.00
		st	84.33	0.78	71.25	98.95	91.43	67.50	73.33
	KNN	stg	85.00	0.79	71.25	95.79	97.14	67.50	80.00
		s	85.33	0.80	73.75	97.89	91.43	67.50	86.67
		sg	84.00	0.78	78.75	94.74	91.43	60.00	73.33
		st	84.66	0.79	71.25	94.74	98.57	65.00	80.00
	NB	stg	77.66	0.70	53.75	87.37	98.57	65.00	80.00
		s	79.33	0.72	53.75	92.63	95.71	70.00	80.00
		sg	81.33	0.75	58.75	93.68	97.14	72.50	73.33
		st	76.00	0.68	50.00	88.42	98.57	57.50	80.00
	RF	stg	87.00	0.82	71.25	97.89	98.57	75.00	80.00
		s	85.33	0.80	71.25	72.50	92.86	72.50	80.00
		sg	84.66	0.79	72.50	96.84	94.29	67.50	73.33
		st	85.66	0.80	71.25	97.89	98.57	65.00	80.00
	SVM	stg	86.66	0.82	76.25	96.84	91.43	75.00	86.67
		s	82.00	0.75	71.25	97.89	82.86	62.50	100.00
		sg	85.00	0.79	78.75	97.89	87.14	62.50	86.67
		st	85.33	0.80	73.75	96.84	95.71	62.50	86.67
Área	Algoritmos	Variáveis	Acurácia						
			Global (%)	Kappa	Vegetação (%)	Pastagem (%)	Café (%)	Outros usos (%)	Água (%)
III	DT	stg	85.66	0.80	78.57	96.00	92.00	62.50	86.67
		s	84.00	0.78	74.29	96.00	91.00	62.50	80.00
		sg	84.00	0.78	74.29	96.00	91.00	62.50	80.00
		st	85.66	0.80	78.57	96.00	92.00	62.50	86.67
	KNN	stg	85.66	0.80	77.14	93.33	96.00	62.50	80.00
		s	86.00	0.81	77.14	96.00	94.00	65.00	80.00
		sg	82.66	0.76	78.57	93.33	89.00	55.00	92.31
		st	86.00	0.81	78.57	90.67	99.00	60.00	80.00
	NB	stg	82.00	0.75	64.29	90.67	94.00	67.50	80.00
		s	84.00	0.78	65.71	93.33	95.00	75.00	73.33
		sg	84.66	0.79	68.57	96.00	93.00	72.50	80.00
		st	82.00	0.75	60.00	93.33	95.00	65.00	86.67
	RF	stg	86.00	0.81	77.14	98.67	93.00	62.50	80.00
		s	85.00	0.79	74.29	97.33	92.00	65.00	80.00
		sg	85.00	0.79	72.86	98.67	92.00	65.00	80.00
		st	85.00	0.79	72.86	98.67	92.00	65.00	80.00
	SVM	stg	87.66	0.83	78.57	96.00	99.00	62.50	80.00
		s	86.33	0.81	80.00	96.00	96.00	57.50	80.00
		sg	86.33	0.81	81.43	92.00	96.00	62.50	80.00
		st	88.33	0.84	78.57	100.00	99.00	60.00	80.00

3.1 Desempenho dos algoritmos de classificação

As classificações geradas apresentaram desempenhos diferentes, tanto em relação ao algoritmo de classificação usado, quanto pelas variáveis empregadas. O algoritmo SVM apresentou os melhores resultados para as três áreas, sendo que na área II o algoritmo RF apresentou acurácia igual ao algoritmo SVM, conforme Tabela 3. Os piores resultados foram obtidos pelos algoritmos RF na área I e NB nas áreas II e III. Quian et al. (2015), comparando diferentes algoritmos de AM, como o RF e NB, mostraram que o SVM apresentou os melhores desempenhos de classificação, corroborando os resultados encontrados neste estudo. O desempenho do algoritmo SVM está relacionado à escolha dos parâmetros de Kernel, no qual o Kernel Radial Basis Function (RBF) é o mais recomendado pela literatura e apresenta os melhores resultados de acurácia (Pradhan, 2013). Neste estudo, utilizou-se o Kernel RBF para as classificações utilizando SVM e isso pode ter colaborado para que este algoritmo apresentasse os melhores resultados de classificação.

Na área I, os índices de acerto global das classificações variaram entre 76,66% e 85,33%, especificado na Tabela 3. As melhores classificações foram geradas usando o algoritmo SVM, com os conjuntos de varáveis *s* (acurácia global = 85,33% e índice *Kappa* = 0,80) e *st* (acurácia global = 84,66% e índice *Kappa* = 0,80). Os resultados com menor acurácia foram usando o algoritmo RF, com as variáveis *s* (acurácia global = 76,66% e índice *Kappa* = 0,69), *sg* com acurácia global de 77,33% e índice *Kappa* de 0,70, *stg* (acurácia global = 78,33% e índice *Kappa* = 0,71) e usando o algoritmo NB com as variáveis *stg* (acurácia global = 79,66% e índice *Kappa* = 0,73). Todos os outros algoritmos usados, independente do conjunto de variáveis, obtiveram resultados de acurácia global e índice *Kappa* superiores. Resultados distintos foram encontrados por alguns autores (Duro et al., 2012; Gislason et al., 2006; Pal, 2005) os quais mostraram que o RF vem obtendo bons desempenhos nas classificações, inclusive quando comparado a outros algoritmos de classificação (Li et al., 2014; Gislason et al., 2006; Pal, 2005).

Na área II, a porcentagem de acerto global ficou entre 76,00% e 87,00%, de acordo com a Tabela 3. As melhores classificações foram geradas utilizando-se os algoritmos SVM e RF e o mesmo conjunto de variáveis *stg*, com acurácia global de 87,00% e 86,66%, respectivamente. No entanto, o índice *Kappa* foi idêntico nas duas classificações (0,82). O menor desempenho foi obtido pelo algoritmo NB, com os conjuntos de variáveis *st* (acurácia global = 76,00% e índice *Kappa* = 0,69), *stg* (acurácia global = 77,66%) e *s* com 79,33% de acerto. Todas as classificações usando o algoritmo NB, independente das variáveis utilizadas, apresentaram os menores índices de acerto, quando comparadas aos outros algoritmos.

Outros estudos indicaram que ambos os algoritmos RF e SVM podem alcançar resultados de acurácia global semelhantes e que são tipicamente maiores do que aqueles obtidos utilizando algoritmos como DT, como mostrado por Pal (2005), que relatou que tanto SVM quanto RF produziram precisões de classificação semelhantes. Gislason et al. (2006) demonstraram que os modelos baseados RF alcançaram precisões de classificação mais elevadas do que as produzidas por DT. Estes resultados são diferentes dos relatados por Li et al. (2013), usando imagens Landsat TM, em que o classificador NB obteve um desempenho um pouco acima do SVM (86,60% NB e 85,90% SVM), porém, estes autores trabalharam mapeando vegetação aquática.

Amostras mais homogêneas e em grande quantidade podem melhorar o desempenho de classificação do algoritmo NB, tornando-o mais preciso, uma vez que este algoritmo é sensível ao tamanho e à uniformidade das amostras de treinamento, uma vez que utiliza amostras de treinamento para estimar os valores dos parâmetros para a distribuição dos dados (Qian et al., 2015).

Na área III, os índices de acerto global variaram de 82,00% a 88,33%, de acordo com a Tabela 3. As melhores classificações obtidas foram geradas utilizando-se o classificador SVM, para todos os conjuntos de variáveis usadas. O índice de acerto mais alto foi alcançado usando as variáveis *st* com acurácia global de 88,33% e índice *Kappa* de 0,84, seguido de *stg* (acurácia global = 87,66% e índice *Kappa* = 0,83). Os menores índices de acerto foram obtidos usando o algoritmo NB, com as variáveis *stg* e *st*, ambas com acerto de 82,00%. Resultados melhores foram obtidos utilizando-se o algoritmo DT com as variáveis *sg* e *s*, todas com o mesmo índice de acerto (acurácia global = 84,00% e índice *Kappa* = 0,78). Li et al. (2014) também encontraram resultados semelhantes, realizando o mapeamento de uso da terra numa região da China. Eles mostraram que os classificadores SVM e RF apresentaram bons resultados na separabilidade entre as classes. O algoritmo DT pode mostrar desvantagens, pois a árvore pode conter muitas ramificações, o que torna a interpretação da classificação difícil (Hussain et al., 2013).

Na área II, foi verificada a maior variação entre as acurácias (11,00%), provavelmente pelo fato de esta área apresentar uma configuração fisiográfica bem diferente das demais. Esta região é muito íngreme, com relevo bastante acidentado, apresentando algumas áreas sombreadas na imagem, o que pode levar interpretações distintas entre os algoritmos de classificação (Andrade et al., 2013a). Já na área I, a variação foi menor (8,67%), porém, as diferenças nas classificações podem ser explicadas pela configuração da paisagem. As maiores confusões foram encontradas na classe outros usos, que foi bastante confundida com a classe pastagem em todos os algoritmos usados. A área III foi a que apresentou menor amplitude entre as acurácias (6,33%), além de mostrar os maiores índices de acerto. Este resultado era esperado devido ao fato de a região apresentar uma estrutura fisiográfica relativamente menos complexa, com áreas mais homogêneas e com relevo suave ondulado predominante. Estes fatores podem contribuir para o bom desempenho na classificação (Bertoldo, 2008).

Alguns trabalhos usando algoritmos de AM mostraram que o desempenho destes classificadores tem sido superior, quando comparado ao dos demais algoritmos de classificação de imagens (Li et al., 2014; Pradhan, 2013; Otukei & Blaschke, 2010). Entretanto, alguns algoritmos de AM apresentam melhores resultados que outros, como é o caso do SVM e RF (Wieland et al., 2014; Li et al., 2014; Duro et al., 2012; Gislason et al. 2006). Wieland et al. (2014) apresentaram resultados promissores no uso de diferentes algoritmos de AM para a classificação em diferentes sensores, no qual os algoritmos SVM e RF apresentaram os melhores desempenhos de classificação, e NB com os piores resultados. Estes resultados corroboram com os obtidos neste estudo, no qual as melhores acurácias foram obtidas pelos algoritmos SVM e RF, nas áreas II e III.

Analisando-se as comparações entre as melhores classificações, conforme Tabela 4, observa-se que os algoritmos utilizados não interferem na acurácia, visto que não houve diferenças significativas na comparação entre os mesmos. Na área I é possível observar que as maiores diferenças foram em relação ao algoritmo RF; na área II, as maiores diferenças envolvem o algoritmo NB e, na área III, as maiores diferenças envolvem o algoritmo SVM.

Tabela 4: Matriz de comparações entre as melhores classificações de cada algoritmo de AM. As comparações foram realizadas pelo teste de McNemar com chi-quadrado tabulado X^2=3,81; a) área I; b) área II; c) área III.

a)	DT	KNN	NB	RF	SVM
DT	*				
KNN	0	*			
NB	0.06	0	*		
RF	0.43	0.54	0.81	*	
SVM	0.24	0.41	0.06	1.31	*

b)	DT	KNN	NB	RF	SVM
DT	*				
KNN	0.02	*			
NB	0.54	0.80	*		
RF	0.42	0.24	1.92	*	
SVM	0.32	0.16	1.70	0	*

c)	DT	KNN	NB	RF	SVM
DT	*				
KNN	0.08	*			
NB	0.06	0.10	*		
RF	0	0	0.10	*	
SVM	0.42	0.32	0.80	0.32	*

Mesmo não apresentando diferença significativa entre as melhores classificações, é possível observar que as classificações apresentam diferenças entre si, como ilustrado na Figura 3. Na área I, as classificações apresentaram a mesma acurácia (índice *Kappa* = 0,75; acurácia global 81,33%) usando o algoritmo DT, com os atributos *stg*, *sg* e *st*, além das classificações utilizando o conjunto de variáveis *s* e *sg* (índice *Kappa* = 0,76; acurácia global 81,66%), usando o algoritmo KNN. No entanto, algumas áreas na imagem foram classificadas de forma diferente. Isto pode ser observado também na área II, com os quatro conjuntos de classificação (índice *Kappa* = 0,74; acurácia global 84,33%), usando o algoritmo DT e na área III isso ocorre para, pelo menos, dois conjuntos de variáveis em cada classificador usado. Estes resultados mostram que, mesmo possuindo o mesmo índice de acerto, as classificações são diferentes, conforme Figura 3. Portanto, além das análises estatísticas e de acurácia, é fundamental uma interpretação visual das classificações, para selecionar a que melhor separou as classes de uso.

Algumas classificações mostraram diferenças significativas, quando comparadas com o mesmo conjunto de variáveis, porém, diferenciando o algoritmo usado, mostrado na Tabela 5. Estes resultados foram observados em duas áreas (I e II), nas quais as classificações obtiveram a maior amplitude entre as acurácias. Na área I, as únicas classificações que mostraram diferença significativa foram aquelas usando variáveis espectrais, entre os algoritmos RF e SVM, pois elas mostraram grande diferença na acurácia, de acordo com Tabela 3. Nesta área, o algoritmo RF foi o que apresentou as menores acurácias, independente do conjunto de variáveis usados. A área II foi a que mostrou maior número de diferenças significativas entre as classificações. As classificações usando as variáveis *stg* somente apresentaram diferenças significativas utilizando o algoritmo NB, comparadas com RF e com SVM, mostrado na Tabela 5. Comparando o algoritmo NB com os demais, utilizando as variáveis *st*, todas as classificações foram diferentes, significativamente. As demais comparações entre algoritmos não mostraram diferença significativas, entretanto, é possível observar que as maiores variações ocorrem quando se compara NB aos demais classificadores. Todas as classificações na área III não mostraram diferenças significativas. Resultados semelhantes foram encontrados por Duro et al. (2012) e Quian et al. (2015), que também não verificaram diferenças significativas usando DT, KNN, RF e SVM.

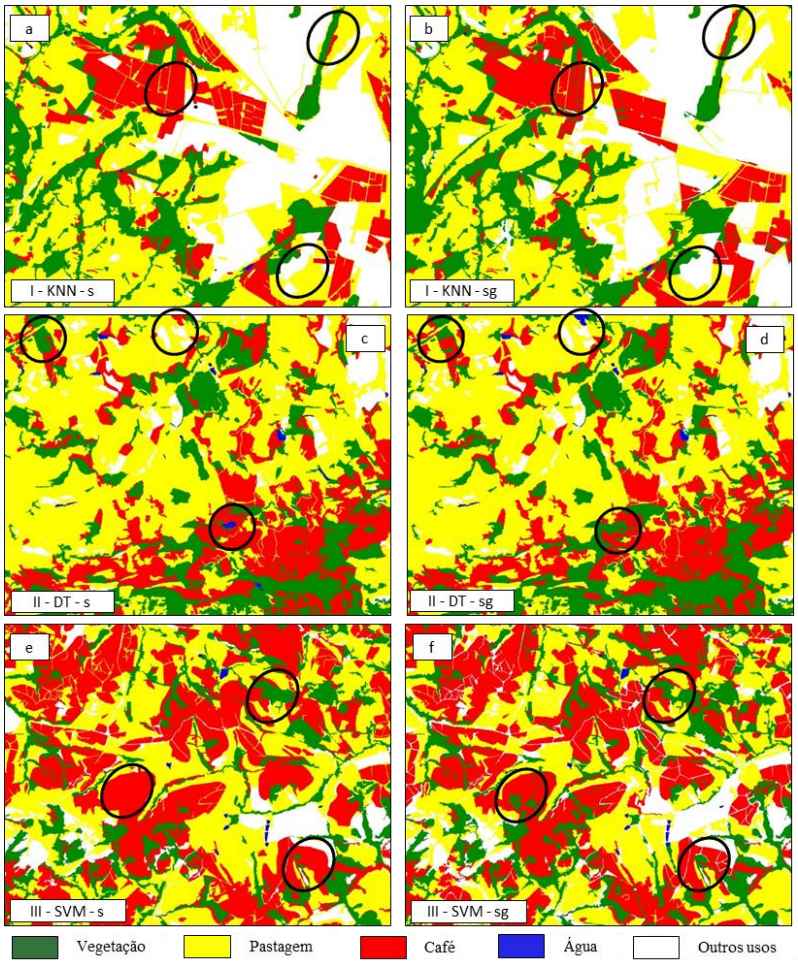

Figura 3: Diferença entre classificações com o mesmo índice de acerto. a) Área I, algoritmo KNN e variável *s* (s); b) Área I, algoritmo KNN e variável *sg*; c) Área II, algoritmo DT e variável *s*; d) Área II, algoritmo DT e variável *sg*; e) Área III, algoritmo SVMe variável *s*; f) Área III, algoritmo SVMe variável *sg*.

Estes resultados colaboram para afirmar que o algoritmo RF, na área I, e NB, nas áreas II e III, apresentaram os resultados mais distintos entre as classificações. Os demais classificadores mostraram um comportamento mutável nas três áreas, em determinados momentos mostrando melhores desempenhos, com índices de acurácias maiores, e, em outros, piores, com índices de acurácia mais baixos. Já o algoritmo SVM foi o mais eficiente por apresentar resultados acurados para as três áreas analisadas, mesmo usando diferentes conjuntos de variáveis, portanto, sendo recomendado para a classificação de áreas cafeeiras.

Tabela 5: Matriz de comparações usando o mesmo conjunto de variáveis, porém, diferenciando o algoritmo usado. As comparações foram realizadas pelo teste McNemar com chi-quadrado tabulado X^2=3,81. Em negrito, o que foi significativo; a) área I; b) área II; c) área III.

a)	DT-stg	KNN-stg	NB-stg	RF-stg	SVM-stg
DT-stg	*				
KNN-stg	0.16	*			
NB-stg	0.16	0.66	*		
RF-stg	0.54	1.30	0.10	*	
SVM-stg	0.10	0.00	0.54	1.12	*

a)	DT-s	KNN-s	NB-s	RF-s	SVM-s
DT-s	*				
KNN-s	0.06	*			
NB-s	0.06	0.24	*		
RF-s	2.16	1.50	2.94	*	
SVM-s	0.42	0.80	0.16	**4.50**	*

a)	DT-st	KNN-st	NB-st	RF-st	SVM-st
DT-st	*				
KNN-st	0.06	*			
NB-st	0.00	0.06	*		
RF-st	0.10	0.32	0.10	*	
SVM-st	0.66	0.32	0.16	1.30	*

a)	DT-sg	KNN-sg	NB-sg	RF-sg	SVM-sg
DT-sg	*				
KNN-sg	0.00	*			
NB-sg	0.16	0.10	*		
RF-sg	0.96	1.12	1.92	*	
SVM-sg	0.06	0.42	0.42	0.54	*

b)	DT-stg	KNN-stg	NB-stg	RF-stg	SVM-stg
DT-stg	*				
KNN-stg	0.02	*			
NB-stg	2.66	3.22	*		
RF-stg	0.42	0.24	**5.22**	*	
SVM-stg	0.32	0.32	**4.86**	0.00	*

b)	DT-s	KNN-s	NB-s	RF-s	SVM-s
DT-s	*				
KNN-s	0.06	*			
NB-s	1.50	2.16	*		
RF-s	0.06	0.00	2.16	*	
SVM-s	0.32	0.66	0.42	0.66	*

b)	DT-st	KNN-st	NB-st	RF-st	SVM-st
DT-st	*				
KNN-st	0.00	*			
NB-st	**4.16**	**4.50**	*		
RF-st	0.10	0.06	**5.60**	*	
SVM-st	0.06	0.02	**5.22**	0.00	*

b)	DT-sg	KNN-sg	NB-sg	RF-sg	SVM-sg
DT-sg	*				
KNN-sg	0.00	*			
NB-sg	0.54	0.42	*		
RF-sg	0.00	0.02	0.66	*	
SVM-sg	0.02	0.06	0.80	0.00	*

c)	DT-stg	KNN-stg	NB-stg	RF-stg	SVM-stg
DT-stg	*				
KNN-stg	0.00	*			
NB-stg	0.80	0.80	*		
RF-stg	0.00	0.00	0.96	*	
SVM-stg	0.24	0.24	1.92	0.16	*

c)	DT-s	KNN-s	NB-s	RF-s	SVM-s
DT-s	*				
KNN-s	0.24	*			
NB-s	0.00	0.24	*		
RF-s	0.06	0.06	0.06	*	
SVM-s	0.32	0.00	0.32	0.10	*

c)	DT-st	KNN-st	NB-st	RF-st	SVM-st
DT-st	*				
KNN-st	0.00	*			
NB-st	0.80	0.96	*		
RF-st	0.06	0.10	0.42	*	
SVM-st	0.42	0.32	2.40	0.80	*

c)	DT-sg	KNN-sg	NB-sg	RF-sg	SVM-sg
DT-sg	*				
KNN-sg	0.10	*			
NB-sg	0.02	0.24	*		
RF-sg	0.06	0.32	0.00	*	
SVM-sg	0.32	0.80	0.16	0.10	*

3.2 Variáveis mais eficientes para separação das classes

Um total de 14 características espectrais, 25 texturais e 11 geométricas foi utilizado como variáveis para as classificações. Estes conjuntos foram testados individualmente e em conjunto para identificar qual a contribuição e a eficiência destas variáveis no processo de classificação. As classificações geradas utilizando somente variáveis texturais e geométricas apresentaram índices de acerto muito baixos, valores discrepantes em relação às classificações geradas utilizando variáveis espectrais. Todas as classificações geradas usando o conjunto de variáveis

espectrais apresentaram bons índices de acurácia, acima de 75% de acerto, de acordo com Tabela 3, para as três áreas. Todas as classificações mostram que as características espectrais são as mais importantes, seguidas por características texturais, enquanto as características geométricas mostram a menor importância para a separabilidade das classes, como ilustrado na Figura 4.

As variáveis espectrais podem discriminar melhor as diferenças no comportamento entre os alvos da superfície terrestre (Araujo & Mello, 2010), enquanto as variáveis de textura e geometria são secundárias, auxiliando no processo de classificação.

Isso demonstra como os dados espectrais foram fundamentais para uma boa acurácia nas classificações, em todas as áreas estudadas. Resultados semelhantes foram encontrados por Wieland et al. (2014), estudando o comportamento destas variáveis para a classificação de áreas urbanas, em que os dados espectrais se destacaram pela melhor separabilidade das classes, seguidos das variáveis de textura.

Na área I, as maiores acurácias das classificações, foram utilizando o conjunto de variáveis *s*, com os algoritmos DT, NB, e SVM, porém usando as mesmas variáveis (*s*), usando o algoritmo KNN, mostrado na Tabela 3. Utilizando as variáveis texturais, os melhores resultados foram obtidos usando os classificadores RF e KNN. Os resultados na área II mostraram que melhores classificações foram geradas utilizando-se o conjunto de variáveis *stg* e *st*, em todos os classificadores utilizados. Estes resultados também foram percebidos na área III, em que as melhores classificações foram oriundas do conjunto de variáveis que continham dados texturais para todos os classificadores, exceto para o classificador NB, cuja melhor classificação foi com o conjunto de variáveis *sg*. Nas três áreas, os melhores resultados foram obtidos utilizando-se os conjuntos de variáveis *stg* ou *st*, para todos os algoritmos testados, mostrando que variáveis texturais podem ser importantes na classificação.

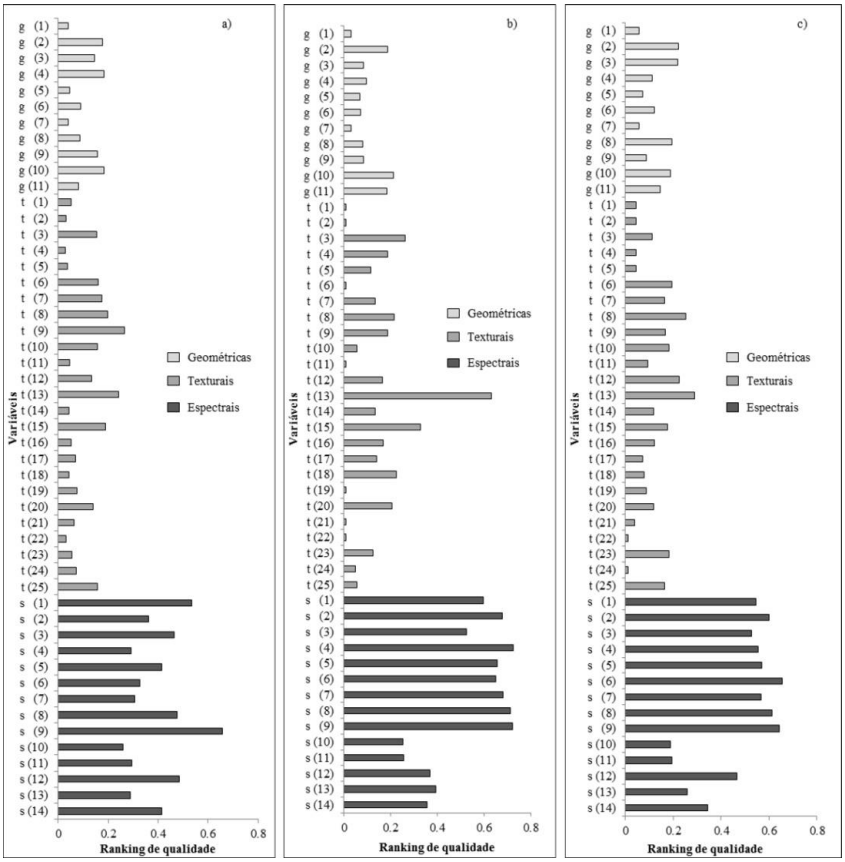

Figura 4: Pontuação das variáveis usadas no processo de classificação, obtidas com os algoritmos RF e DT. a) I; b) II; c) III. Variáveis: s (espectrais); t (texturais) e g (geométricas).

Para as três áreas, em relação aos conjuntos de variáveis, não houve diferença estatisticamente significativa entre as classificações, independente do algoritmo de classificação utilizado, de acordo com a Tabela 6. As diferenças significativas foram observadas somente quando não foram utilizadas variáveis espectrais nas classificações. As características de textura podem auxiliar muito no processo de classificação de imagens de Sensoriamento Remoto, principalmente quando se lida com áreas de grande heterogeneidade espectral (Ruiz et al., 2004). Segundo os mesmos autores, as características intrínsecas dos diversos objetos na superfície terrestre podem apresentar uma boa alternativa para distinguir as diferentes classes de uso. Porém, como foi observado neste estudo, para melhorar a classificação, estas variáveis precisam estar associadas a variáveis espectrais.

Tabela 6: Matriz de comparações usando o mesmo algoritmo de classificação, porém, diferenciando o conjunto de variável usada. As comparações foram realizadas pelo teste de McNemar com chi-quadrado tabulado X^2=3,81; a) área I; b) área II; c) área III.

a)

DT	stg	s	st	sg
stg	*			
s	0.11	*		
st	0.00	0.11	*	
sg	0.00	0.11	0.00	*

KNN	stg	s	st	sg
stg	*			
s	0.11	*		
st	0.03	0	*	
sg	0.11	0.03	0.03	*

NB	stg	s	st	sg
stg	*			
s	0.96	*		
st	0.17	0.33	*	
sg	0.67	0.03	0.17	*

RF	stg	s	st	sg
stg	*			
s	0.17	*		
st	0.17	0.67	*	
sg	0.06	0.03	0.43	*

SVM	stg	s	st	sg
stg	*			
s	0.43	*		
st	0.24	0.03	*	
sg	0.33	1.5	1.13	*

b)

DT	stg	s	st	sg
stg	*			
s	0.00	*		
st	0.00	0.00	*	
sg	0.00	0.00	0.00	*

KNN	stg	s	st	sg
stg	*			
s	0.00	*		
st	0.00	0.02	*	
sg	0.06	0.10	0.02	*

NB	stg	s	st	sg
stg	*			
s	0.16	*		
st	0.16	0.66	*	
sg	0.80	0.24	1.70	*

RF	stg	s	st	sg
stg	*			
s	0.16	*		
st	0.10	0.00	*	
sg	0.32	0.02	0.06	*

SVM	stg	s	st	sg
stg	*			
s	1.30	*		
st	0.10	0.66	*	
sg	0.16	0.54	0.00	*

c)

DT	stg	s	st	sg
stg	*			
s	0.16	*		
st	0.00	0.16	*	
sg	0.16	0.00	0.16	*

KNN	stg	s	st	sg
stg	*			
s	0.00	*		
st	0.00	0.00	*	
sg	0.54	0.66	0.66	*

NB	stg	s	st	sg
stg	*			
s	0.24	*		
st	0.00	0.24	*	
sg	0.42	0.02	0.42	*

RF	stg	s	st	sg
stg	*			
s	0.06	*		
st	0.10	0.00	*	
sg	0.06	0.00	0.00	*

SVM	stg	s	st	sg
stg	*			
s	0.10	*		
st	0.02	0.24	*	
sg	0.10	0.00	0.24	*

Já as variáveis geométricas não foram muito eficazes na separação entre classes, sobretudo entre objetos como café, floresta e pastagem, principalmente porque os objetos destas classes, muitas vezes, apresentarem praticamente o mesmo formato, tamanho, assimetria, etc., o que pode ter dificultado uma melhor separação pelos algoritmos. Quando foram usados juntamente com variáveis espectrais (*sg*), a acurácia melhorou, porém, a maioria dos resultados obtidos por este conjunto de variáveis ficou abaixo dos índices de acerto de outros conjuntos, como *s*, *st* e *stg*. O uso de múltiplas medidas de forma permite uma melhor discriminação entre os objetos e melhora a classificação de imagens (Van der Werff &Van der Meer, 2008), porém, estes autores trabalharam com áreas espectralmente semelhantes, o que tornou as variáveis geométricas importantes no processo de classificação, diferentemente deste estudo, no qual se trabalhou com áreas espectralmente heterogêneas. Além disso, segundo Witten et al. (2011), as características redundantes ou irrelevantes que fornecem pouca informação para uma classificação específica podem ter um efeito negativo sobre modelos de AM e podem levar a uma diminuição na acurácia da classificação. Portanto, as variáveis de geometria podem não ter desempenhado um papel tão importante no processo de separação das classes de cobertura da terra.

Durante a fase de treinamento de um classificador, os próprios algoritmos de AM já selecionam as variáveis importantes e ignoraram as irrelevantes ou redundantes (Witten et al., 2011). Como visto na Figura 3, as classificações geradas nas três áreas mostraram que as variáveis mais expressivas foram as espectrais, seguidas das texturais e, por último, as geométricas, em todos os conjuntos de variáveis utilizados (*s*, *st*, *sg* e *stg*).

Todas as áreas mostraram os índices de vegetação NDVI e SAVI como uma das variáveis mais importantes para separabilidade nas classificações. Estes índices estão entre os principais utilizados no mapeamento da cobertura da terra (Machado et al., 2010). Diferentemente dos resultados obtidos por Sarmiento et al. (2014), a data de aquisição das imagens pode ter influenciado positivamente a capacidade dos índices em separar as classes de cobertura da terra, uma vez que, embora estas apresentem uma resposta espectral semelhante, a aquisição das imagens no período seco contribuiu para que as classes café e pastagem apresentassem níveis de biomassa verde distintos e, consequentemente, apresentando valores elevados de NDVI e SAVI.

As 10 primeiras variáveis mais significativas na área I e III foram todas espectrais; já na área II, as 10 melhores foram, na grande maioria, espectrais, porém, a variável GLCM Homogeneidade b5 (banda do infravermelho próximo) foi importante no processo de classificação. A Homogeneidade fornece uma medida da distribuição dos valores de intensidade dos pixels e quanto maior for o valor dado por esta métrica, maior será similaridade dos pixels (Ruiz, et al., 2004). De acordo com os mesmos autores, as técnicas de textura são muito eficientes na classificação de paisagens que contenham uma elevada heterogeneidade espectral, isto ocorre na área II, uma vez que a área possui uma grande variabilidade espectral. Isso mostra a importância destas variáveis no processo de classificação usando imagens de Sensoriamento Remoto. Segundo Souza et al. (2009), as variáveis de textura descrevem padrões de suavidade, rugosidade e regularidade dos alvos, sendo características importantes para reconhecer e classificar objetos.

É possível observar que, em todas as áreas, as maiores diferenças foram geradas quando se utilizaram variáveis texturais em algum momento das classificações, exceto na área II, usando KNN, na qual a melhor classificação foi obtida usando o conjunto de variáveis *s*. Nesse sentido, os resultados mostraram que as variáveis texturais, quando combinadas com as espectrais, podem trazer alguns benefícios à separabilidade das classes, porém, não representam diferença estatisticamente significante. Ruiz et al. (2004) também atestaram isso em um estudo realizado em áreas florestais e urbanas, no qual as variáveis de textura forneceram uma alternativa para auxiliar as variáveis espectrais para a classificação de unidades florestais com uma alta

heterogeneidade espectral, ou quando as classes são definidas pelas diferenças na densidade da vegetação nativa.

Além disso, é importante ressaltar que muitas comparações não apresentaram nenhuma diferença estatística. Assim, é possível afirmar que, independente do conjunto de variáveis usadas, é imprescindível que as variáveis espectrais estejam presentes no processo de classificação de imagens de Sensoriamento Remoto.

3.3 Melhores variáveis para separação da classe café

Nas três áreas, os conjuntos de variáveis apresentaram performances diferentes para a classificação do café, para todos os algoritmos utilizados. O maior índice de acerto, na acurácia do produtor, foi verificado para a área III, com 99,00% e o menor para a área II, com índice acerto de 82,86%, como pode ser visto na Tabela 3.

O maior índice de acerto obtido na área I foi verificado utilizando-se o conjunto *s* (98,46%), usando o algoritmo NB, enquanto o menor acerto foi usando a variável *sg*, usando o algoritmo RF, com 87,69% de acurácia, mostrado na Tabela 3. A maioria dos conjuntos de variáveis *s* foi os que geraram melhores índices de acerto em todos os algoritmos, porém, é importante salientar que quanto maior foi o índice de acerto desta classe, menores foram os índices de acerto para a classe vegetação nativa, sendo confundida com a classe café. O algoritmo RF usando o conjunto de variáveis *s* apresentou bastante confusão para as áreas de vegetação nativa, que foram classificadas, em sua maioria, como café. Isto pode ter ocorrido devido ao fato de a área ser composta por cerrado, o que pode ter causado uma confusão espectral maior com áreas cafeeiras. Todos os outros conjuntos de variáveis obtiveram índices variados, variando o percentual de acerto para a classe café.

Na área II, o conjunto de variáveis *st* e *stg* foi o que mostrou melhores índices de acerto para a classe café (98,57%), em todos os classificadores testados, enquanto os menores índices foram obtidos utilizando-se a variável *s*, seguido da *sg*, conforme Tabela 3. A classe vegetação nativa também foi confundida com a classe café, mostrando índices de acerto bem baixos quando usado o algoritmo NB para todos os conjuntos de variáveis empregados.

A área III foi a que apresentou os maiores índices de acerto para o café; todas as classificações obtiveram índices acima de 89,00%. Os menores índices foram obtidos utilizando o conjunto de variáveis *sg*, em todos os classificadores usados. Nesta área também foi verificado que a classe vegetação nativa foi confundida com a classe café, obtendo índices mais baixos de acerto.

Outros estudos realizados mapeando café indicaram resultados semelhantes, no qual áreas de café foram bem classificadas e áreas de vegetação nativa apresentaram índices de acerto mais baixos (Martínez-Verduzco et al., 2012; Andrade et al., 2013b; Sarmiento et al., 2014).

É possível observar, em sua maioria, as melhores variáveis para separar o café foram aquelas que continham dados texturais, porém, sempre estando associadas aos dados espectrais. Alguns trabalhos mostram que variáveis texturais podem auxiliar na melhoria da qualidade do mapeamento em áreas cafeeiras (Marujo et al., 2013; Santos et al., 2012), assim como alguns índices de vegetação, como o NDVI (Cordero-Sancho & Sader, 2007). Este fato também ocorreu neste estudo, avaliando áreas distintas, o que mostra que as variáveis texturais podem favorecer a classificação de áreas cafeeiras. Neste caso, a melhoria na qualidade das classificações pode ser decorrente dos algoritmos utilizados e não das variáveis usadas, uma vez que a maioria dos resultados obtidos utilizando algoritmos de AM produziu bons índices de acerto.

A classe café foi bem classificada nas três áreas, para acurácia do produtor, porém, algumas áreas de vegetação nativa foram classificadas como café, apresentando índices de acerto menores, principalmente nas áreas II e III, de acordo com a Tabela 3. Na área I, os maiores erros na classe vegetação nativa foram encontrados para o classificador RF, para todos os conjuntos de variáveis utilizados. Para as áreas II e III, os índices mais baixos foram obtidos utilizando o classificador NB, também independente do conjunto de variáveis empregadas. Analisando a acurácia do usuário, observou-se que a classe café apresentou índices mais baixos, sendo confundida com a classe vegetação nativa. Estes resultados mostram que, mesmo usando um conjunto de variáveis mais robusto e diferentes algoritmos para classificação, a vegetação e o café ainda são confundidos no processo de classificação. Adami et al. (2009) e Moreira et al. (2004) afirmam que o mapeamento de áreas cafeeiras, muitas vezes, é uma tarefa difícil, pois há uma grande confusão espectral entre os diferentes tipos de cobertura com respostas espectrais semelhantes, como é o caso da vegetação nativa e do café. De acordo com Li et al. (2014), o desempenho dos algoritmos pode ser pior em áreas mais complexas, em paisagens mais dinâmicas, devido à heterogeneidade espectral existente.

4. CONCLUSÕES

Neste trabalho avaliaram-se a eficácia do uso de diferentes algoritmos de aprendizagem de máquina e a importância de diferentes conjuntos de variáveis para o mapeamento da cafeicultura, em três áreas ambientalmente distintas. De acordo com os resultados apresentados, pode-se chegar às seguintes conclusões:

1) nas áreas I e II, houve diferença significativa entre algumas classificações que utilizaram o mesmo conjunto de variáveis, porém, diferenciando o algoritmo usado. Entretanto, para todas as áreas, as classificações que utilizaram diferentes combinações de variáveis não mostraram diferença significativa entre elas;

2) a qualidade das classificações não apresentou diferença estatisticamente significativa. No presente trabalho, os mapas produzidos a partir da combinação de variáveis espectrais e texturais resultaram em valores numericamente superiores para a qualidade de discriminação entre as classes café e vegetação nativa;

3) os algoritmos mais eficientes para classificar cafezais, no presente estudo, foram SVM e RF, na área II e SVM, nas áreas I e III. No entanto, o algoritmo SVM foi o mais robusto, apresentando os melhores resultados para todas as áreas analisadas, usando diferentes conjuntos de variáveis;

4) durante o processo de classificação, a classe mais confundida com a classe café foi a classe vegetação nativa. As áreas de pastagem não apresentaram confusão com as áreas de café;

5) apesar das confusões no processo de classificação, as informações espectrais são fundamentais para obter uma acurácia mais elevada.

Recomendam-se mais testes para melhorar ainda mais a separabilidade entre as classes vegetação e café, uma vez que foram muito confundidas. Mais estudos estão sendo realizados para avaliar atributos temporais e outros sensores para melhorar o mapeamento de café em Minas Gerais.

REFERÊNCIAS BIBLIOGRÁFICAS

Adami, Marcos, Mauricio Alves Moreira, Marco Aurélio Barros, Bernardo Friedrich, and Theodor Rudorff. "Avaliação Da Exatidão Do Mapeamento Da Cultura Do Café No Estado de Minas Gerais." In *XIV Simpósio Brasileiro de Sensoriamento Remoto*, 1–8. 2009.

Andrade, Alexandre Curvelo, Cristiane Nunes Francisco, and Cláudia Maria de Almeida. "DESEMPENHO DE CLASSIFICADORES PARAMÉTRICO E NÃO PARAMÉTRICO NA CLASSIFICAÇÃO DA FISIONOMIA VEGETAL Evaluating the Performance of Parametric and Non-Parametric Classi Fi Ers for Identifying Vegetal Physiognomies Universidade Federal Fluminense – UFF INTROD." *Revista Brasileira de Cartografia* 65 (2): 227–41. 2013.

Andrade, Lívia Naiara, Tatiana Grossi Chquiloff Vieira, Wilian Soares Lacerda, Margarete Marin Lordelo Volpato, and Clodoveu Augusto Davis Junior. "APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA CLASSIFICAÇÃO DE ÁREAS CAFEEIRAS EM MACHADO - MG." *Coffee Science* 8 (1): 78–90. 2013.

Araujo, Thiago Peixoto de, and Fernando Machado de Mello. "Processamento de Imagens Digitais - Raz??o Entre Bandas." *Geociencias* 29 (1): 121–31. 2010.

Bertoldo, Mathilde Aparecida. "Caracterização Edafoambiental Da Cefeicultura Na Região de Três Pontas, Minas Gerais." *Thesis*, Universidade Federal de Lavras, 2008.

CONAB. "Acompanhamento de Safra Brasileira." 2014.

Cordero- Sancho, S., and S. a. Sader. "Spectral Analysis and Classification Accuracy of Coffee Crops Using Landsat and a Topographic- environmental Model." *International Journal of Remote Sensing* 28 (7): 1577–93. 2007. doi:10.1080/01431160600887680.

Duro, Dennis C., Steven E. Franklin, and Monique G. Dubé. "A Comparison of Pixel-Based and Object-Based Image Analysis with Selected Machine Learning Algorithms for the Classification of Agricultural Landscapes Using SPOT-5 HRG Imagery." *Remote Sensing of Environment* 118 (March). Elsevier Inc.: 259–72. 2012.doi:10.1016/j.rse.2011.11.020.

Foody, Giles M. "Thematic Map Comparison : Evaluating the Statistical Significance of Differences in Classification Accuracy." *Photogrammetric Engineering & Remote Sensing* 70 (5): 627–33. 2004.

Gislason, Pall Oskar, Jon Atli Benediktsson, and Johannes R. Sveinsson. "Random Forests for Land Cover Classification." *Pattern Recognition Letters* 27 (4): 294–300. 2006.doi:10.1016/j.patrec.2005.08.011.

Gomez, C., M. Mangeas, M. Petit, C. Corbane, P. Hamon, S. Hamon, a. De Kochko, D. Le Pierres, V. Poncet, and M. Despinoy. "Use of High-Resolution Satellite Imagery in an Integrated Model to Predict the Distribution of Shade Coffee Tree Hybrid Zones." *Remote Sensing of Environment* 114 (11). Elsevier Inc.: 2731–44. 2010.doi:10.1016/j.rse.2010.06.007.

Haralick, Robert M., K. Shanmugam, and Its'Hak Dinstein. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (6): 610–21. doi:10.1109/TSMC.1973.4309314. 1973.

Hussain, Masroor, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. "Change Detection from Remotely Sensed Images: From Pixel-Based to Object-Based Approaches." *ISPRS Journal of Photogrammetry and Remote Sensing* 80 (June). International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS): 91–106. 2013.

doi:10.1016/j.isprsjprs.2013.03.006.

Landis, J Richard, and Gary G Koch. “The Measurement of Observer Agreement for Categorical Data Data for Categorical of Observer Agreement The Measurement.” *Biometrics* 33 (1): 159–74. 1977.

Li, Congcong, Jie Wang, Lei Wang, Luanyun Hu, and Peng Gong. “Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery.” *Remote Sensing* 6 (2): 964–83. 2014. doi:10.3390/rs6020964.

Machado, M. L., H. M. R. Alves, T. G. C. Vieira, E. I Fernades-Filho, and M. P. C. Lacerda. “Mapeamento de Áreas Cafeeiras (Coffea Arabica L.) Da Zona Da Mata Mineira Usando Sensoriamento Remoto.” *Coffee Science* 5 (2): 113–22. 2010.

Martínez-Verduzco, Guillermo C., J. Mauricio Galeana-Pizaña, and Gustavo M. Cruz-Bello. “Coupling Community Mapping and Supervised Classification to Discriminate Shade Coffee from Natural Vegetation.” *Applied Geography* 34 (May). Elsevier Ltd: 1–9. 2012. doi:10.1016/j.apgeog.2011.10.001.

Marujo, R. F. B., M. M. L. Volpato, T. G. C. Vieira, H. M. R. Alves, and M. B. P. Ribeiro. “Classificação Orientada a Objetos Aplicada a Cultivos Cafeeiros Em Três Pontas - MG.” In *XVI Simpósio Brasileiro de Sensoriamento Remoto*, 1338–45. 2013.

Moreira, Mauricio Alves, Marcos Adami, and Friedrich Theodor. “Análise Espectral E Temporal Da Cultura Do Café Em Imagens Landsat Spectral and Temporal Behavior Analysis of Coffee Crop in Landsat Images.” *Pesquisa Agropecuária Brasileira* 39 (3): 223–31. 2004.

Niel, T Van, T Mcvicar, and B Datt. “On the Relationship between Training Sample Size and Data Dimensionality: Monte Carlo Analysis of Broadband Multi-Temporal Classification.” *Remote Sensing of Environment* 98 (4): 468–80. 2005. doi:10.1016/j.rse.2005.08.011.

Organização Internacional De Café - OIC. "Promoção e desenvolvimento de mercado"**.** 1 p. (OIC). Disponível em: http://www.ico.org/. Acesso em: 25/11/2014

Otukei, J.R., and T. Blaschke. “Land Cover Change Assessment Using Decision Trees, Support Vector Machines and Maximum Likelihood Classification Algorithms.” *International Journal of Applied Earth Observation and Geoinformation* 12 (February): S27–31. 2010. doi:10.1016/j.jag.2009.11.002.

Pal, M. “Random Forest Classifier for Remote Sensing Classification.” *International Journal of Remote Sensing* 26 (1): 217–22. 2005. doi:10.1080/01431160412331269698.

Pradhan, Biswajeet. “A Comparative Study on the Predictive Ability of the Decision Tree, Support Vector Machine and Neuro-Fuzzy Models in Landslide Susceptibility Mapping Using GIS.” *Computers & Geosciences* 51 (February). Elsevier: 350–65. 2013. doi:10.1016/j.cageo.2012.08.023.

Ruiz, L A, A Fdez-Sarría, and J A Recio. “TEXTURE FEATURE EXTRACTION FOR CLASSIFICATION OF REMOTE SENSING DATA USING WAVELET DECOMPOSITION : A COMPARATIVE STUDY.” In *20th ISPRS Congress*, 1–6. 2004.

Santos, Jefersson Alex dos, Philippe-Henri Gosselin, Sylvie Philipp-Foliguet, Ricardo S Torres, and Alexandre Xavier Falcão. “Multiscale Classification of Remote Sensing Images.” *IEE Transactions on Geoscience and Remote Sensing* 50 (10): 3764–75. 2012.

Sarmiento, Christiany Mattioli, Gláucia Miranda Ramirez, Priscila Pereira Coltri, Luis Felipe Lima Silva, Otávio Augusto Carvalho Nassur, and Jefferson Francisco Soares. “Comparação de classifiCadores Supervisionados Na Discriminação de Áreas Cafeeiras Em Campos Gerais -

Minas Gerais." *Coffee Science* 9 (4): 546–57. 2014.

Souza, Vanessa Cristina Oliveira, Tatiana Grossi Chquiloff Vieira, Margarete Marin Lordelo Volpato, and Helena Maria Ramos Alves. "Espacialização E Dinâmica Da Cafeicultura Mineira Entre 1990 E 2008, Utilizando Técnicas de Geoprocessamento." *Coffee Science* 7 (2): 122–34. 2012.

Werff, H van der. M. A., and F. D. van der Meer. "Shape-Based Classification of Spectrally Identical Objects." *ISPRS Journal of Photogrammetry and Remote Sensing* 63 (2): 251–58. 2008. doi:10.1016/j.isprsjprs.2007.09.007.

Velloso, Marcos Henrique, "Coffee inventory through orbital imagery". Rio de Janeiro: *Instituto Brasileiro do Café*, 20p. (SR-525). 1974.

Vieira, Tatiana Grossi Chquiloff, Helena Maria Ramos Alves, Mathilde Aparecida Bertoldo, and Vanessa Cristina Oliveira de Souza. "GEOTHECNOLOGIES IN THE ASSESSMENT OF LAND USE CHANGES." *Coffee Science* 2 (2): 142–49. 2007.

Wieland, Marc, and Massimiliano Pittore. "Performance Evaluation of Machine Learning Algorithms for Urban Pattern Recognition from Multi-Spectral Satellite Images." *Remote Sensing* 6 (4): 2912–39. 2014. doi:10.3390/rs6042912.

Witten, Ian. H.; Frank, Eibe.; Hall, Mark. A. "Data mining: Practical machine learning tools and techniques", 3rd ed., p. 629. 2011.

Recebido em maio de 2015.
Aceito em abril de 2016.