# Seleção de Co-Variáveis para Classificação de Imagens de Satélite Através do Algoritmo Random Forest

Ligia Faria Tavares de Souza [1]
Leiliane Bozzi Zeferino [1]
Elpídio Inácio Fernandes Filho [1]
Liovando Marciano da Costa [1]
Ariecha Vieira Rodrigues Tibiriçá [1]

[1] Universidade Federal de Viçosa – UFV
Departamento de Solos – Avenida Peter Henry Rolfs, s/n, Campus Universitário – 36570-900 – Viçosa, MG, Brasil
ligiaftsouza@gmail.com
leilibz@gmail.com
elpidio@ufv.br
liovandomc@yahoo.com.br
ariechavrt@gmail.com

**Abstract.** This papers aims to present a method to select co-variables using the Random Forest classifier to classify satellite images, by the land use classification of Lagoa Formosa, a county in Minas Gerais state. The data used was Landsat 8 images for the dry and wet seasons, cartography data of the county obtained from IBGE, such as geomorphology, vegetation, boundary, localities and roads at the 1:250.000 scale, and a geologic map of the area at 1:100.000 scale and a soils map at 1:250.00 scale. The SRTM was used to obtain the digital elevation model and other topography co-variables. With these in hand, 98 co-variables were obtained, being 26 spectral co-variables, 45 topography co-variables, 5 geomorphological co-variables, 2 geological co-variables, 3 soils co-variables, 3 vegetation co-variables and 14 Euclidian distance co-variables. The numeric co-variables were analyzed by non-linear correlation, and the categorical co-variables were analyzed by dissimilarity to eliminate those statistically similar. From the 58 co-variables analyzed by the Random Forest classifier, a list with the models with 20 to 5 co-variables was generated with the Kappa index. The selected model was that with the highest value for the Kappa index, the one with 16 co-variables, to obtain the simplest model that better explain the final classification of land use. However, it is still necessary to develop other techniques that help decrease the researcher subjectivity to choose the final model.

**Palavras-chave:** remote sensing, data mining, numeric and categorical data, sensoriamento remoto, mineração de dados, dados numéricos e categóricos.

## 1. Introdução

As técnicas de Sensoriamento Remoto têm como objetivo extrair informações das imagens digitais através da atribuição de significado a um pixel pelas suas propriedades numéricas, e esse processo é chamado genericamente de “classificação” (Novo, 2010). De acordo com Meneses e Sano (2012), as técnicas de classificação digital de imagens automatizam esse processo de extração das informações, eliminando a subjetividade da interpretação humana, o que reduz o esforço do analista.

A mineração de dados (em inglês, data mining) é o processo de descoberta de padrões em dados (KDD, *Knowledge Discovery in Databases*), o que engloba uma série de técnicas e algoritmos utilizados para a construção de um modelo de conhecimento, que pode ser resultado de um processo automático ou semiautomático (Witten & Frank, 2011; Vieira et al., 2012). Esses padrões são significativos, pois apresentam uma forma de predizer tendências e comportamentos futuros, auxiliando na tomada de decisão do usuário baseado no conhecimento obtido (Boulila et al., 2011; Witten & Frank, 2011). Segundo Rogan et al. (2008), processos de mineração de dados podem ser representados também como árvores de classificação. Assim, o classificador *Random Forest* é um tipo de tecnologia de mineração de dados (Naidoo et al., 2012).

O classificador *Random Forest*, termo geral para designar método ensemble (que combina o uso de classificadores) que utiliza classificadores do tipo árvore de decisão, é a combinação de árvores de decisão onde cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores. Assim, consiste em uma coleção de classificadores estruturados em árvores, onde o arquivo de saída da classificação é dado pela maioria dos votos dados pelas árvores (Breiman, 2001).

Dentre as vantagens do *Random Forest*, pode-se listar a boa acurácia obtida, mesmo quando comparada a processos como o *Adaboost*, podendo até mesmo ser melhor; a robustez do classificador quanto aos *outliers* (valores discrepantes) e aos ruídos; processamento mais rápido do que de métodos como *bagging*; além de ser um método simples e de apresentar informações importantes como a estimativa de erro interna, a correlação e a importância das co-variáveis (Breiman, 2001). No entanto, esse algoritmo é também considerado como uma caixa-preta, já que pouco se sabe sobre como sua assinatura estatística é processada, pois não é possível separar e analisar cada árvore de decisão individualmente (Gislason et al., 2006; Naidoo et al., 2012; Grinand et al., 2013).

O objetivo desse trabalho é apresentar uma metodologia para escolha de co-variáveis com o uso do algoritmo *Random Forest* para classificar imagens de satélite, através da classificação do uso e ocupação do município de Lagoa Formosa, em Minas Gerais. Dessa forma, a classificação final poderá seguir um raciocínio lógico estatístico, diminuindo a interferência da subjetividade do pesquisador.

## 2. Materiais e Métodos

### 2.1. Área de Estudo

A área de estudo é o município de Lagoa Formosa, que está localizado no centro-oeste do estado de Minas Gerais, a aproximadamente 378 km de distância de Belo Horizonte, e 27 km de Patos de Minas, como mostra a Figura 1. Possui população estimada em 18.037 habitantes (IBGE, 2016) e sua economia é baseada na agropecuária, comércio e prestação de serviços, destacando-se como maior produtor de feijão da região (Câmara Municipal de Lagoa Formosa, 2016).

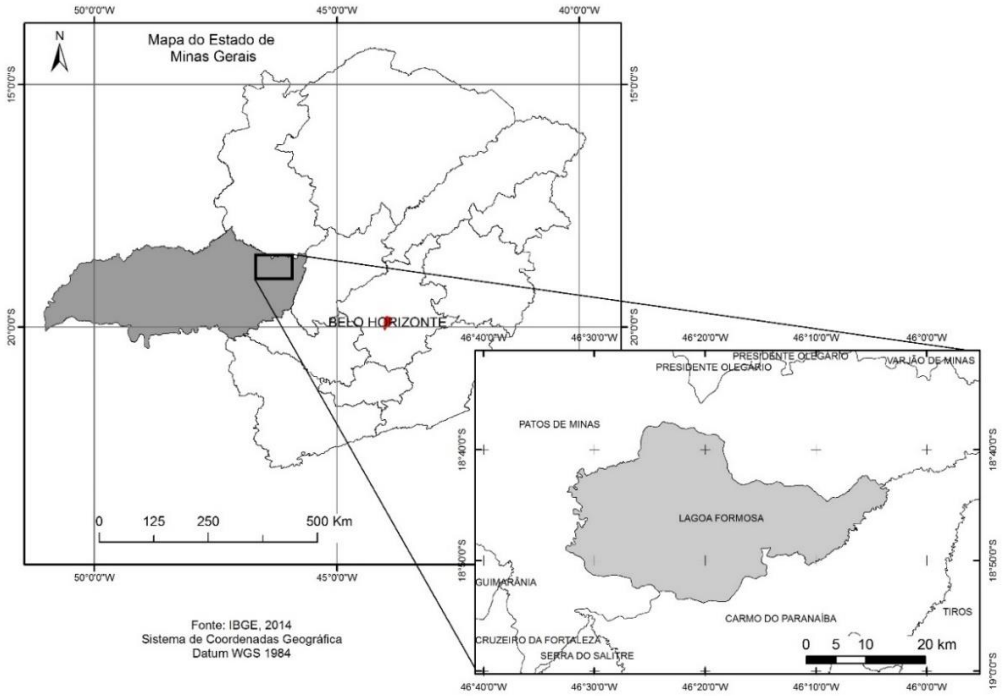


Figura 1. Mapa de localização da área de estudo no estado de Minas Gerais.

## 2.2. Bases de Dados

Para esse trabalho, as imagens utilizadas são do Landsat-8, disponíveis no catálogo da USGS (United States Geological Survey), e correspondem à época seca e à época chuvosa do município de Lagoa Formosa, no estado de Minas Gerais, conforme mostra a Tabela 1.

Tabela 1. Imagens Landsat-8 obtidas no catálogo da USGS.

| Imagem | Data | Órbita/Ponto |
|---|---|---|
| LC82200732014126LGN00 | 06/05/2014 | 220/73 |
| LC82190732014135LGN00 | 15/05/2014 | 219/73 |
| LC82200732015289LGN00 | 16/10/2015 | 220/73 |
| LC82190732015282LGN00 | 09/10/2015 | 219/73 |

Dados cartográficos do município de Lagoa Formosa, em Minas Gerais, como limite municipal, estradas e localidades (igrejas, escolas e outros), na escala de 1:100.000, foram obtidos junto ao IBGE, assim como os dados de geomorfologia e vegetação, na escala de 1:250.000. Os dados de geologia foram obtidos pelo mapeamento realizado pelo Projeto Alto Paranaíba, realizado pela CODEMIG. A folha Carmo do Paranaíba foi utilizada, e está na escala de 1:100.000 (Uhlein et al., 2011). O mapeamento pedológico utilizado foi realizado pela Embrapa (Motta et al., 2004), na escala de 1:250.000. Foram utilizadas imagens SRTM (Shuttle Radar Topographic Mission), da NASA, para retirar o modelo de elevação e gerar outras co-variáveis de terreno.

As diferenças na escala dessas informações são dadas pela escassez de dados para a região de estudo. No entanto, acredita-se que esses dados possam auxiliar na melhor compreensão do ambiente, ainda que as escalas não sejam iguais.

O processamento dos dados foi realizado no software ArcGIS® 10.1, do ESRI (Environmental Systems Research Institute). Além desse, foi utilizado o R (v. 3.2.4), programa gratuito disponibilizado na internet baseado em linguagem computacional para manipulação de dados e análises estatísticas (R Core Team, 2016).

## 2.3. Amostras de Treinamento e Validação

As amostras para treinamento e validação das imagens foram coletadas com auxílio do Google Earth® e do Basemap do ArcGIS 10.1. Foram coletados 250 polígonos de 12 pixels cada da imagem Landsat 8, de forma aleatória. Cada pixel coletado equivale a um ponto, correspondendo a 3000 pontos amostrais.

## 2.4. Co-Variáveis Preditivas

Com as imagens de satélite e a base cartográfica reunida, foram geradas 98 co-variáveis da região, divididas em: 26 espectrais, 45 topográficas, 5 geomorfológicas, 2 geológicas, 3 pedológicas, 3 de vegetação e 14 de distância euclidiana de alguns dados de interesse.

As co-variáveis espectrais foram obtidas através das imagens Landsat-8 tanto para a época seca quanto para a época chuvosa. Além das bandas 2, 3, 4, 5, 6 e 7, foi feita uma imagem composta com essas bandas, e também foram gerados seis (6) índices espectrais, sendo eles: NDVI (*Normalized Difference Vegetation Index*), NDSI (*Normalized Difference Soil Index*), SAVI *(Soil Adjusted Vegetation Index*), *Clay Minerals* (minerais de argila) e *Iron Oxides* (óxidos de ferro). As equações dos índices espectrais estão na Tabela 2. O SAVI foi desenvolvido como melhoria do NDVI, pois aplica uma constante L que minimiza os efeitos da cor do solo nos seus resultados. A constante está no intervalo entre 0 e 1. O valor de L=1 é utilizado quando a densidade da vegetação é baixa; L=0,5 quando a densidade da vegetação é média; e L=0,25 quando a densidade da vegetação é alta (Huete, 1988). Para esse estudo, foram

gerados dois índices SAVI para cada época, um com fator de correção para densidade baixa e outro com fator de correção para densidade média, ou seja, foram utilizados os valores de 1 e 0,5.

Tabela 2. Equações dos índices espectrais gerados.

| **Índices Espectrais** | | |
|---|---|---|
| $NDVI = \frac{B5 - B4}{B5 + B4}$ | $SAVI = \frac{(B5 - B4)(1 + L)}{B5 + B4 + L}$ | $IRON\ OXIDES = \frac{B4}{B2}$ |
| $NDSI = \frac{B7 - B3}{B7 + B3}$ | $CLAY\ MINERALS = \frac{B6}{B7}$ | |

As co-variáveis topográficas foram derivadas do modelo digital de elevação (MDE) obtido com a imagem SRTM. No ambiente R, o MDE foi inserido a fim de gerar outras 44 co-variáveis com o pacote RSAGA. No final foram totalizadas 45 co-variáveis topográficas, incluindo o modelo digital de elevação (MDE).

Com os dados geomorfológicos disponibilizados pelo IBGE, foi possível utilizar informações como a forma do modelado geomorfológico, nome dos domínios morfoestruturais, nome da região do modelado, nome da unidade geomorfológica, além da geomorfologia geral. Cada um desses foi inserido e analisado separadamente.

Das co-variáveis geológicas, foram utilizadas a litologia da região e as unidades geomorfológicas, que também foram analisadas separadamente.

Das informações sobre a pedologia, foram utilizadas separadamente a classificação de solos no primeiro nível categórico, a classificação até o terceiro nível categórico, e a classificação até o quarto nível categórico dos solos, podendo incluir associações de solos, de acordo com o mapeamento de Motta et al. (2004). Ainda que contenham informações muito similares, a área de ocorrência muda com a quantidade de classes apresentadas. A classificação até o quarto nível categórico, por exemplo, mostra a espacialização de 16 tipos de solos diferentes para a área de estudo, enquanto que a classificação até o segundo nível categórico é mais concisa e resumida.

A vegetação disponibilizada pelo IBGE apresentou alguns dados relevantes que foram incluídos nas análises para o mapeamento. As informações utilizadas foram a de vegetação (ou área natural) ou antropismo do principal componente, a de vegetações dominantes e subdominantes, e o tipo de cobertura vegetal. Nesse caso, a lógica para o uso de informações semelhantes é a mesma utilizada para a compreensão das informações de solos similares, pois ainda que sejam informações parecidas e que estejam contidas umas nas outras, não se sabe qual a escala que a vegetação dada pelo IBGE pode afetar no mapeamento final.

A distância euclidiana é uma medida da similaridade entre dois objetos, sendo a distância matemática entre dois pontos, e ao usar a distância como medida de proximidade, menores distâncias indicam maior similaridade (Hair et al., 2009). Dessa forma, a distância euclidiana é uma medida que pode auxiliar na compreensão da espacialização dos usos do solo, pois pode apresentar padrões que afetam a ocupação no município, como a distância da rede de drenagem e das estradas, tanto urbanas quanto rurais.

### 2.5. Seleção das Co-Variáveis pelo Random Forest

As amostras coletadas foram inseridas no ambiente R junto às 98 co-variáveis desejáveis para análise, para que, através da interface desse software com o ArcGIS 10.1®, as informações fossem associadas aos pontos de interesse.

Foram obtidas a moda e mediana dos pontos de cada polígono das amostras coletadas. Segundo Costa Neto (1977), a moda é uma medida de posição, pois indica a região das máximas frequências, ou seja, de um dado conjunto de valores, a moda corresponde ao valor (ou valores)

de máxima frequência. De acordo com o mesmo autor, a mediana, que também é uma medida de posição, é uma quantidade que busca caracterizar o centro da distribuição de frequências. Dessa forma, as amostras apresentam o valor mais significativo das respostas espectrais ao contrário de apresentar valores pontuais que podem não ser representativos para a área.

Das 98 co-variáveis inseridas para análise, foram eliminadas aquelas que apresentaram variância zero, além de separar as co-variáveis categóricas, como geologia, solos, vegetação e geomorfologia, das co-variáveis contínuas, ou numéricas, compostas pelas co-variáveis espectrais, topográficas e das distâncias euclidianas. A inserção de todas essas co-variáveis para análise no programa R teve como objetivo deixar que o algoritmo selecionasse aquelas que achasse mais relevante para a classificação.

As co-variáveis contínuas foram submetidas à análise de correlação não linear para eliminar co-variáveis estatisticamente similares, sendo esse método mais robusto do que a correlação linear, mais comumente utilizada. Essa correlação não linear foi feita de acordo com o Coeficiente de Dependência Aleatorizado (RDC – Randomized Dependence Correlation) proposto por Lopez-Paz et al. (2013). O RDC é uma medida não linear de dependência entre amostras aleatórias multivariadas, que se baseia em correlações canônicas de variáveis aleatórias e em cópulas. Correlação canônica é um procedimento estatístico multivariado que funciona como uma extensão de uma regressão múltipla, onde uma co-variável é explicada por uma combinação linear de outras variáveis. Porém, nessa correlação não existe distinção entre variáveis dependentes e independentes, buscando-se somente a máxima correlação entre ambas (James & McCulloch, 1990; Hair et al., 2009). As cópulas, por sua vez, permitem construir um modelo multivariado capaz de separar o comportamento marginal das variáveis aleatórias da estrutura de dependência que pode existir entre elas (Li, 2000; Poczos, Ghahramani, Schneider, 2012).

Já as co-variáveis categóricas foram submetidas à análise de dissimilaridade, que é um método de análise de agrupamento, sendo que quanto maior for a sua medida, menor será a semelhança entre os indivíduos (Hair et al, 2009). A técnica utilizada para análise simultânea das co-variáveis contínuas e das co-variáveis categóricas foi aquela proposta por Gower (1971), onde esses dados qualitativos são submetidos a uma matriz de distância apresentando valores entre 0 e 1, com o valor 1 correspondendo à máxima diferença. Essa metodologia, até o momento, não tem sido amplamente aplicada para mapeamentos, pois não há relatos que mostram a distinção entre co-variáveis categóricas e co-variáveis contínuas. Essas co-variáveis têm sido submetidas à análise de correlação, o que revela a presença de um erro grosseiro durante a garimpagem das co-variáveis preditivas. O uso da dissimilaridade tem sido mais utilizado na área de genética (Rodriguez et al., 2005; Vieira et al, 2007; Rocha et al., 2010).

As co-variáveis restantes foram então submetidas à seleção de importância feita pelo algoritmo Random Forest, estabelecendo um limite máximo de 20 co-variáveis para que fosse escolhido um modelo suficientemente elucidativo e simples para a classificação da imagem. Das co-variáveis escolhidas, foi gerada uma lista com as co-variáveis dispostas pela sua importância na classificação, acompanhado com os respectivos índices Kappa, fator importante para auxiliar na escolha do modelo classificado.

## 3. Resultados e Discussão

Com a eliminação das co-variáveis de variação zero, foram excluídas seis co-variáveis, todas topográficas. Dessa forma, foram eliminadas as co-variáveis que não apresentaram variação dos dados para nenhum dos pontos amostrais.

Pela correlação não-linear, as 79 co-variáveis contínuas foram analisadas e 28 dessas mostraram correlação maior que 95%, sendo então excluídas do conjunto total. Foram 14 co-variáveis espectrais, 12 co-variáveis topográficas e 2 co-variáveis de distância euclidiana foram excluídas, restando 51 co-variáveis contínuas. Depois de testes realizados com o ponto de corte

da correlação, o valor de 95% foi estabelecido por ter eliminado uma quantidade significativa de co-variáveis sem prejudicar o modelo final, o qual foi aferido através da comparação dos valores de índice Kappa antes e depois da análise de correlação.

Do total de 13 co-variáveis categóricas, a dissimilaridade eliminou 6 co-variáveis com limiar de 0,5 (50%). Foram eliminadas 4 co-variáveis geomorfológicas, 1 co-variável geológica e 1 co-variável de vegetação. O valor mínimo de corte da dissimilaridade foi estabelecido de forma semelhante ao valor da correlação não linear. Valores menores ou maiores que 50% para corte por dissimilaridade afetaram a escolha dos fatores importantes para a classificação final.

A seleção das co-variáveis pela importância no *Random Forest* foi feita com 58 co-variáveis. O total de co-variáveis contínuas analisadas foi de 51, sendo 12 espectrais, 27 de terreno e 12 de distância euclidiana. Foram analisadas 7 co-variáveis categóricas, sendo 1 geomorfológica, 1 geológica, 3 de vegetação e 2 pedológicas. Como resultado, o algoritmo ranqueou as co-variáveis escolhidas de acordo com a sua importância para a classificação da imagem, com tamanho máximo de 20 co-variáveis para o modelo. Estabelecendo esse limite, a intenção é escolher um modelo com o menor número de co-variáveis possíveis que melhor representem o ambiente de estudo, caracterizando um modelo parcimonioso.

O modelo selecionado apresenta 16 co-variáveis e possui o melhor índice Kappa entre os modelos de 20 até 5 co-variáveis. No entanto, a diferença desses índices entre os diferentes modelos pode ser muito pequena, como se observa no Gráfico 1. Como a diferença no índice Kappa entre o modelo com 16 e o de 12 co-variáveis é menor que 0,1, e a intenção é um modelo mais simples, com menos co-variáveis, procedeu-se a uma análise de comparação das áreas classificadas com o mesmo uso. Foi possível notar as diferenças quando comparados os modelos com 16 co-variáveis e os modelos com menor número de co-variáveis. Considerando o modelo de 16 co-variáveis com 100% de acerto, assumindo que seja o melhor modelo para a área, quando comparado com o modelo de 12 co-variáveis, esse apresenta índice de acertos de 91%, ainda que a diferença com o índice Kappa seja menor que 0,1 e o modelo ainda seja apontado como excelente.

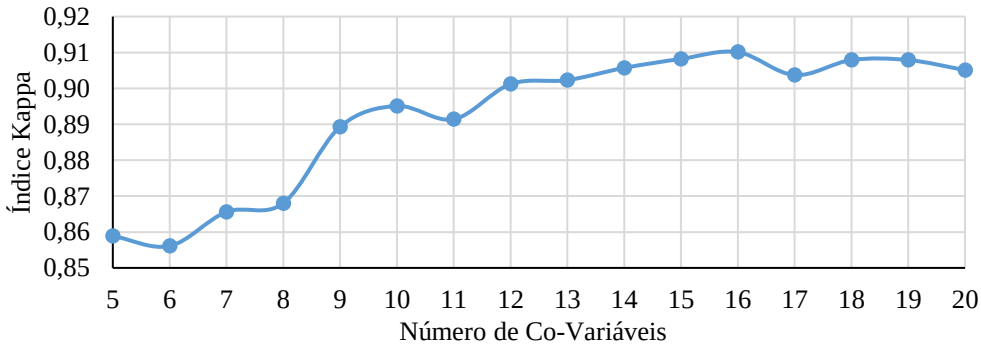


Gráfico 1: Variação no Índice Kappa de acordo com o número de co-variáveis do modelo.

Obviamente que a comparação entre as classificações não é um dado muito preciso, afinal dessa forma assume-se que uma classificação está correta. No entanto, essa comparação nos dá uma ideia das áreas de maior erro nas classificações, além de auxiliar no processo de coleta de amostras extras para melhorar o resultado final.

A escolha do modelo com 16 co-variáveis se deu pela análise visual das classificações em comparação com as imagens de satélite e por ter apresentado o melhor índice Kappa. Dessas co-variáveis, 10 são espectrais, 2 são de distância euclidiana, 3 são topográficas, e 1 é pedológica.

## 4. Conclusões

O principal objetivo desse trabalho foi apresentar uma nova metodologia de seleção de co-variáveis com o algoritmo Random Forest. As co-variáveis inseridas diferiram entre numéricas e categóricas, pois unindo-as espera-se que as classificações sejam melhores.

A seleção das co-variáveis é feita de forma automática, utilizando-se de métodos estatísticos para a seleção final daquelas mais importantes. Essa seleção final depende do conjunto inicial de co-variáveis a serem analisadas, bem como das suas características, além da quantidade de co-variáveis desejadas para o modelo.

Os valores de correlação e dissimilaridade apresentados nesse trabalho foram estabelecidos com base nos valores de índice Kappa. Esses valores irão depender do conjunto de co-variáveis trabalhadas, assim como a realidade local do ambiente em questão.

Dos conjuntos finais apresentados pelo classificador, ainda é necessário o desenvolvimento de técnicas que diminuam a subjetividade do pesquisador em avaliar se a classificação corresponde bem à realidade da área de estudo.

## Referências Bibliográficas

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

Câmara Municipal de Lagoa Formosa. Lagoa Formosa. Disponível em: <http://camaralagoa.mg.gov.br/cidade>. Acesso em: 20 ago. 2016.

Costa Neto, P. L. de O. **Estatística**. São Paulo: Edgard Blücher, 1977.

Gislason, P. O.; Benediktsson, J. A.; Sveinsson, J. R. Random forest classification of multisource remote sensing and geographic data. Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. **Proceedings**. 2004 IEEE International, v. 2, p. 1049-1052, 2004.

Gower, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, p. 857-871, 1971.

Grinand, C.; Rakotomalala, F.; Gond, V.; Vaudry, R.; Bernoux, M.; Vieilledent, G. Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier. **Remote Sensing of Environment**, v. 139, p. 68-80, 2013.

Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E.; Tatham, R. L. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

Huete, A. R. A soil-adjusted vegetation index (SAVI). **Remote sensing of environment**, v. 25, n. 3, p. 295-309, 1988.

Instituto Brasileiro de Geografia e Estatística (IBGE). Cidades: Minas Gerais, Lagoa Formosa. Disponível em: <http://www.cidades.ibge.gov.br/xtras/perfil.php?lang=&codmun=313750>. Acesso em: 03 out. 2016

James, F. C.; Mcculloch, C. E. Multivariate analysis in ecology and systematics: panacea or Pandora's box?. **Annual Review of Ecology and Systematics**, p. 129-166, 1990.

Li, D. X. On default correlation: A copula function approach. **Journal of Fixed Income**, p. 43-54, 2000.

Lopez-Paz, D.; Hennig, P.; Schölkopf, B. The randomized dependence coefficient. In: A**dvances in Neural Information Processing Systems**, p. 1-9, 2013.

Meneses, P. R.; Sano, E. E. Classificação Pixel A Pixel de Imagens. In: Meneses, P. R.; Almeida, T. de. (Org). **Introdução ao processamento de imagens de sensoriamento remoto**. Brasília: UNB/CNPq, 2012.

Motta, P. E. F. da; Baruqui, A. M.; Santos, H. G. dos. **Levantamento de reconhecimento de média intensidade dos solos da região do alto Paranaíba, Minas Gerais.** Rio de Janeiro: Embrapa Solos, 2004.

Naidoo, L.; Cho, M. A.; Mathieu, R.; Asner, G. Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. **ISPRS Journal of Photogrammetry and Remote Sensing**, v.69, p. 167-179, 2012.

Novo, E. M. L. De M. **Sensoriamento remoto: princípios e aplicações**. 4. ed. São Paulo: Blücher, 2010.
Poczos, B.; Ghahramani, Z.; Schneider, J. Copula-based kernel dependency measures. **Proceedings of the 29 th International Conference on Machine Learning**, Edinburgh, Scotland, UK, 2012.

R Core Team**. R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Viena, Áustria, 2016. Disponível em: <<https://www.r-project.org/>>. Acesso em: 24 jul. 2016

Rocha, M. C.; Gonçalves, L. S. A.; Rodrigues, R.; Silva, P. D.; Carmo, M. D.; Abboud, A. C. D. S. Uso do algoritmo de Gower na determinação da divergência genética entre acessos de tomateiro do grupo cereja**. Acta Scientiarum Agronomy**, v. 32, n. 3, p. 423-431, 2010.

Rodríguez, V. M.; Cartea, M. E.; Padilla, G.; Velasco, P.; Ordás, A. The nabicol: A horticultural crop in northwestern Spain. **Euphytica**, v. 142, n. 3, p. 237-246, 2005.

Rogan, J.; Franklin, J.; Stow, D.; Miller, J.; Woodcock, C.; Roberts, D. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. **Remote Sensing of Environment**, v. 112, n. 5, p. 2272-2283, 2008.

Uhlein, A.; Freitas, A. de M.; Cruz, A. B. da; Silva, Q. F. da; Caxito, F. de A.; Moreira, G. de C. Folha Carmo do Paranaíba SE.23-Y-B-IV. In: Soares, A. C. P.; Noce, C. M.; Fragoso, D. G. C.; Voll, E.; Reis, H. L. S.; Kuchenbecker, M. (eds). **Projeto Alto Paranaíba**. Belo Horizonte: CODEMIG, UFMG, 2011.

Vieira, E. A.; Carvalho, F. I. F.; Bertan, I.; Kopp, M. M.; Zimmer, P. D.; Benin, G.; Silva, J. A. G.; Hartwing, I.; Malone, G.; Oliveira, A. C. Association between genetic distances in wheat (Triticum aestivum L.) as estimated by AFLP and morphological markers. **Genetics and Molecular Biology**, v. 30, n. 2, p. 392-399, 2007

Vieira, M. A.; Formaggio, A. R.; Rennó, C. D.; Atzberger, C.; Aguiar, D. A.; Mello, M. P. Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. **Remote Sensing of Environment**, v. 123, p. 553-562, 2012.

Witten, I. H; Frank, E. **Data Mining: Pratical Machine Learning Tools and Techniques**. Morgan Kaufmann, 2011.