Influência da forma de amostragem na exatidão global e índice kappa

Maola Monique Faria [1]
Elpidio Inácio Fernandes Filho [2]
Márcio Rocha Francelino [2]
Raiza Moniz Faria [2]

[1] Universidade Federal de Roraima - UFRR
Avenida Capitão Ene Garcez, 2413, Aeroporto - 69310-000 – Boa Vista - RR, Brasil
maolageo@gmail.com

[2] Universidade Federal de Viçosa – UFV
Avenida Peter Henry Rolfs, s/n, Campus Universitário, Viçosa – MG, Brasil
{elpidio, marcio.francelino, raiza.faria}@ufv.br

Abstract. The procedure of classifying and grouping pixels of a digital image based on its spectral characteristics using algorithms in a computational program is called image classification. The objective of this article is to evaluate the effect of sampling in the form of polygons and points in global accuracy and in the kappa index in the classification of coffee areas in the Matas de Minas region of the state of Minas Gerais. In addition, the use of cross-validation and validation was evaluated using external data in the kappa index in the classification of coffee areas in the Matas de Minas region of the state of Minas Gerais. A cut of a Landsat 8 scene was used for the area of interest. On this scene, 6,517 polygons were collected, with a mean of 12 pixels, distributed randomly throughout the study area. Based on the samples file in point format, the radiance values of each band of the Landsat 8 image were extracted. Four ways were defined in the definition of training samples of the Random Forest classifier. The procedures were performed using the software interface R and ArcGis 10.2. From the use of randomly collected points, they corroborate the accuracy, global accuracy and kappa, which are higher than those obtained by other treatments when using cross-validation, but the kappa obtained from the external validation is similar to the others.

Palavras-chave: remote sensing, training samples, coffe, sensoriamento remoto, amostras de treinamento, café.

1. Introdução

O procedimento de classificar e agrupar pixels de uma imagem digital com base em suas características espectrais utilizando algoritmos em um programa computacional é denominado classificação de imagens (Richards e Jia, 2006). Esse procedimento pode se dar a partir da interação do analista no treinamento do algoritmo ou não, caracterizando, respectivamente, a classificação supervisionada e a não supervisionada.

O processo de classificação supervisionada requer que o analista treine o algoritmo com base na coleta de amostras das diferentes classes de usos de interesse no estudo em áreas homogêneas, para que com base nessas o classificador possa identificar os pixels espectralmente similares aos das demais amostras. Campbell (1987) recomenda que as amostras de treinamento devem ser coletadas na configuração de polígonos como forma de minimizar a quantidade de vértices nos polígonos utilizados para delimitar a amostra.

Inúmeros são os classificadores supervisionados existentes, sendo que atualmente o classificador Random Forest tem sido apontado pela literatura como sendo um dos mais robusto na separação de classes com similaridade espectral (Caruana; Karampatziakis; Yessenalina, 2008).

Esse classificador foi proposto por Breiman (2001) e seu funcionamento se baseia em uma técnica de agregação de diversos classificadores do tipo árvore de decisão, organizados de forma que sua estrutura seja constituída sempre de forma aleatória (Ghosal; Tikmani; Gupta, 2009). O resultado final da classificação é dado com base na combinação dos resultados das várias árvores de decisão, onde a classificação final é produzida a partir da classe que recebeu o maior número de votos entre todas as árvores existentes no modelo de

predição (Breiman, 2001). Salienta-se que para árvore gerada no modelo é empregado um conjunto de treinamento diferente, formado por n instâncias de treinamento escolhidas aleatoriamente dentro do conjunto de amostras fornecido (Breiman, 2001).

Em relação à avaliação da exatidão dos mapeamentos temáticos, Congalton (1991) descreve os índices kappa e exatidão global. Para a obtenção destes índices emprega-se a matriz de erro, que por sua vez é obtida a partir do cruzamento de dados de validação coletados empregando-se arquivo de amostras independente do que foi utilizado no treinamento do classificador.

Com o advento de novas metodologias e softwares para a realização de classificação de imagens, principalmente as ligadas ao emprego da interface do pacote estatístico R surge a possibilidade da utilização da validação cruzada para a obtenção dos índices de exatidão que anteriormente eram calculados com base em arquivo de validação. O processo de validação cruzada segundo Myers (1997) consiste na remoção de dados do conjunto de dados amostrais e, empregando uma função de estimativa e outra ponderada relacionada com a distância, calcula-se o valor retirado, no caso da classificação de imagens, determina-se o pixel pertencente à determinada classe retirada, usando as amostras remanescentes. Assim, ter-se-á dois valores para o mesmo pixel, o real e o estimado. O erro da estimativa é calculado com base na diferença entre a classe do pixel real e do estimado, sendo repetido para cada pixel amostrado.

Diante disso, o presente trabalho teve por objetivo avaliar o efeito da amostragem na forma de polígonos e de pontos na exatidão global e no índice kappa na classificação das áreas cafeeiras da região das Matas de Minas, estado de Minas Gerais. Além disso, avaliou-se o emprego da validação cruzada e da validação empregando dados externos no índice kappa na classificação das áreas cafeeiras da região das Matas de Minas, estado de Minas Gerais.

2. Metodologia de Trabalho

2.1. Caracterização da área de estudo

A região das Matas de Minas localiza-se na porção sudeste do Estado de Minas Gerais (Figura 1), sendo composta por 63 municípios distribuídos por uma área de 1.749.114 ha, equivalente a 3% da área total do estado de Minas Gerais, com uma população aproximadamente de 900 mil habitantes, que corresponde a 5% da população do Estado (IBGE, 2010). A região produz em média 5 milhões de sacas por ano, que representa 24% da produção do Estado, sendo que 80% das fazendas de café são de até 20 ha, ou seja, de pequenos produtores (IBGE, 2006).

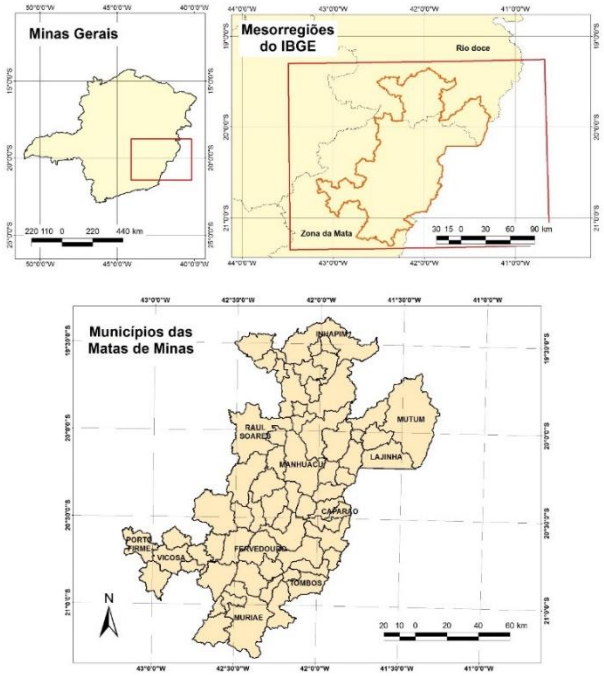

Figura 1. Mapa de localização da região das Matas de Minas.

Caracterizam-se geologicamente por estar localizada sobre rochas datadas do Pré-Cambriano, pertencentes ao Complexo Cristalino, tais como gnaisses diversos e migmatitos, com variado grau de metamorfismo; em menor proporção por metassedimentos, compreendendo quartzitos e mica xistos, e intrusões de basalto; de forma restrita aparecem formações sedimentares recentes, datadas do Terciário e Quaternário (RadamBrasil, 1987; RadamBrasil, 1983).

Sobre essa geologia, desenvolveu-se um relevo com formas diversificadas, onde se destacam áreas planas, onduladas e montanhosas; elevações de topos arredondados com vertentes convexas terminando em vales planos (Valverde, 1958). Ab'Saber (2003) denominou essa paisagem de Domínio dos Mares de Morros.

Segundo Valverde (1958) o clima Cwa é predominante na região e a pluviosidade média anual oscila entre 900 e 1.900 mm.

Originalmente, os municípios eram recobertos pela Mata Atlântica, cuja fitofisionomia constitui o grande conjunto florestal extra-amazônico (Valverde, 1958). Mas, o mapeamento da Flora Nativa e Reflorestamento Regional do Estado de Minas Gerais (Scolforo e Carvalho, 2006), mostra que apenas fragmentos da vegetação nativa são encontrados. Segundo este mapeamento, há na área Floresta Ombrófila, Floresta Semidecídua, Campo e Campo Rupestre e eucalipto como espécie de áreas reflorestadas.

No tocante às classes de solo, predominam os Latossolos e os Argissolos. Nas áreas mais elevadas são encontrados solos com horizonte húmico (UFV, 2010, 2010b).

2.2. Base de dados

O estudo foi realizado utilizando-se as bandas 1 a 7 da imagem, com correção atmosférica, Landsat 8 de 02 de agosto de 2013, disponibilizada no site Earth Explorer (http://earthexplorer.usgs.gov/). Optou-se pela utilização da cena do mês de agosto devido à alta porcentagem de nuvens existentes nas cenas de outras épocas.

A cena Landsat 8 foi recortada de acordo com os limites da área de estudo. Foram utilizados o software ArcGis 10.1 (ESRI, 2011) e o R 3.2 (R Development Core Team, 2015) para o processamento dos dados.

Realizou-se a coleta das amostras de treinamento totalizando um conjunto amostral formado por 6.517 polígonos contendo em média 12 pixels por polígono. Posteriormente, cada polígono foi convertido para raster e em seguida para o formato de pontos totalizando 57.093 pontos amostrais distribuídos de forma aleatória em toda a área de estudo e considerando as oito classes de uso de interesse no estudo: café, mata, eucalipto, solo, água, pastagem, nuvem e sombra. Esses procedimentos foram realizados empregando o software ArcGis 10.1.

Para avaliar o efeito do método de validação sobre os valores de kappa foi selecionado um conjunto aleatório de 1630 polígonos, 25% do total dos polígonos coletados, que foi para a realização da validação externa. Os demais 4887 polígonos foram submetidos a 5 tratamentos

2.3. Tratamentos

Os tratamentos consistiram em 5 esquemas diferentes de separação entre amostras de treinamento e validação. A amostra de trabalho consistiu nos 4.887 polígonos anteriormente separados. Deste conjunto, em todos os tratamentos avaliados, utilizou-se 75% dos dados para treinamento e 25% dos dados para validação, conforme apresentado na Tabela 1. Devido a separação do conjunto de amostras de treinamento e do de validação ser realizado de forma aleatória, foram feitas 10 repetições do processo de classificação, empregando o classificador Random Forest.

Tabela 1. Número de amostras de amostras total, de treinamento e de validação empregada em cada um dos tratamentos avaliados

Tratamento	Número de Amostras total	Amostras de Treinamento	Amostras de Validação
1	42859	32144	10715
2	42859	32160	10699
3	4887	3667	1220
4	4887	3667	1220
5	4887	3667	1220

2.3.1. Tratamento 1

Para a separação das amostras de treinamento e de validação do classificador utilizou-se arquivo de pontos, sendo que durante o processo de separação dos arquivos não foi dado atenção ao fato se estes foram coletados no mesmo polígono de origem. Somente atentou-se para que não houvesse sobreposição entre esles.

2.3.2. Tratamento 2

Para a separação das amostras de treinamento e de validação utilizou-se de polígonos com, em média, 12 pixels cada. Cada polígono foi aleatoriamente identificado como treinamento ou validação e todos os seus pontos foram alocados a uma destas duas classes. Sendo que foi garantido a não sobreposição entre os polígonos de treinamento e os de validação. 75% dos poligonos foram utilizados para treinamento e 25% para validação

2.3.3. Tratamento 3

Neste tratamento, em cada um dos 4887 polígonos foi coletado um ponto de forma aleatória. Após isso, o arquivo de pontos foi separado em arquivo de amostras de treinamento e de validação, onde 75% dos pontos foi utilizado para treinamento e 25% para validação.

2.3.4. Tratamento 4

Neste tratamento para cada um dos 4887 polígonos foi calculada a mediana das co-variáveis, associadas aos pontos existentes em cada polígono. Assim sendo foi gerado um único ponto para cada polígono semelhante ao tratamento 3. Após, empregou-se 75% dos pontos para treinamento e 25% para validação do classificador, garantindo-se a não sobreposição destes.

2.3.5. Tratamento 5

Este tratamento foi semelhante ao tratamento 4 só que foi calculada a média das co-variáveis, associadas aos pontos existentes em cada polígono. Assim sendo foi gerado um único ponto para cada polígono semelhante ao tratamento 3. Após, empregou-se 75% dos pontos para treinamento do classificador e 25% para validação do classificador, garantindo-se a não sobreposição destes.

3. Resultados e Discussão

Na Figura 2 pode-se observar o efeito dos quatro tratamentos no valor dos índices, com base na média das 10 repetições: exatidão global e kappa obtidos a partir da validação cruzada e kappa obtido a partir da validação externa.

O tipo de amostragem adotado no tratamento 1 apesar de ser o com melhor desempenho, este tende a produzir resultados superestimados, isto é, pode ocorrer a coleta do ponto de validação no mesmo polígono que se coletou um ponto de treinamento, apesar de não haver a sobreposição entre os mesmos, existe uma chance muito grande das classes de treinamento e validação serem iguais, por estarem próximas uma da outra e também considerando que cada polígono amostral em geral possui apenas uma classe de uso.

Ao empregar a forma mais utilizada atualmente nos estudos empregando técnicas de classificação de imagens, onde se tem dois conjuntos de polígonos amostrais, um para treinamento e outro para validação. Ao compararmos este aos demais tratamentos aqui avaliados quando se considera o kappa de validação interna (validação cruzada), com a validação externa não se observa diferença entre os valores de kappa

Observou-se que quando do emprego de pontos de treinamento de forma aleatória (tratamento 3) apesar da exatidão global obtida ser considerada excelente, os kappas obtidos são menores que os dos demais tratamentos, demonstrando assim que a forma de amostragem adotada não captou adequadamente a variabilidade espectral das classes existentes, acarretando assim, menor exatidão do mapeamento obtido.

Ao avaliar os índices de avaliação da exatidão obtidos a partir do emprego dos pontos contendo a mediana da resposta espectral de 12 pixels (tratamento 4) observa-se que estes são considerados excelentes pela literatura, porém ao considerarmos a definição da estatística e da teoria da probabilidade, ao empregar a mediana na determinação da resposta espectral das classes de interesse essa pode não estar demonstrando a variabilidade espectral existente dos alvos na natureza, visto que estes apresentam variabilidade em suas assinaturas espectrais advindas de sua estrutura física, química e por conta de fatores externos. Por isso, considerando a definição de média apontada na estatística, a amostragem considerando essa é a mais adequada para fins de classificação de imagens (tratamento 5), visto que será realizado uma média dos valores de resposta espectral da classe de interesse. Isso pode ser comprovado ao observarmos os resultados obtidos quando do emprego do tratamento 5.

Na figura 2 pode-se observar que a validação externa parece ser mais estável em termos dos valores de kappa do que a validação cruzada. Sendo que de modo geral os valores de kappa obtidos na amostragem por polígonos foram maiores do que os obtidos a partir de pontos amostrais.

Ressalta-se que quando do emprego de amostras na forma de polígono, os pixels formadores destes apresentam alta correlação devido à proximidade geográfica. Tal fato pode corroborar para com a obtenção de maiores valores de kappa, visto que ocorrerá superioridade numérica de amostragem. Do ponto de vista computacional a amostragem em polígonos gera amostras de treinamento e validação maiores do que a amostragem por ponto, sendo menos eficientes, em termos de tempo de processamento e espaço de armazenamento.

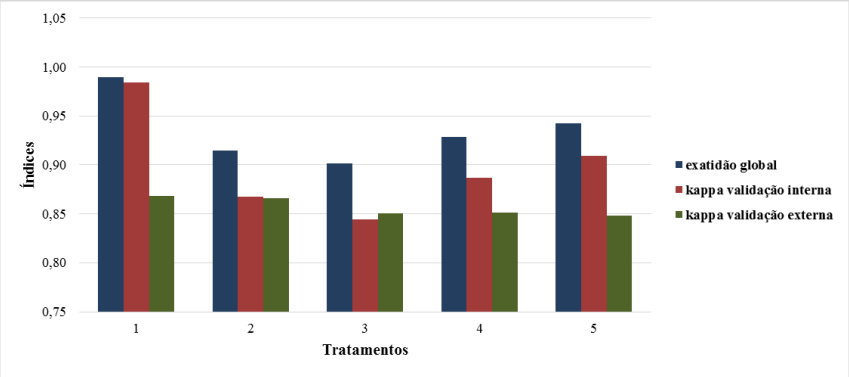

Figura 2. Índices médios obtidos por cada um dos tratamentos avaliados.

4. Conclusões

A amostragem por polígonos impõe que todos os pixels existentes em um polígono sejam alocados na mesma amostra de treinamento ou de validação, de forma mutuamente exclusiva, sob o risco de superestimativa dos valores de kappa.

A separação de um conjunto amostral para validação, apresenta resultados mais consistentes do que a utilização da validação cruzada, para avaliar a exatidão da classificação.

A amostragem por polígonos apresenta valores de kappa ligeiramente maiores que a amostragem por ponto, apesar de ser menos eficiente em termos computacionais.

Agradecimentos

A FAPEMIG pela concessão da bolsa de estudos. Ao Centro de Excelência do Café (CEC), EPAMIG, Embrapa – Café, CNPQ, SEBRAE – MG pelo apoio financeiro.

Referências Bibliográficas

BREIMAN, L. Random Forests. **European Journal of Mathematics**, v. 45, n. 1, p. 5–32, 2001.

CARUANA, R.; KARAMPATZIAKIS, N; YESSENALINA, A. An Empirical evalution of supervised learning in high dimensions. In: International Conference on Machine Learning, 25, Helsinki. **Proceesdings**… Helsinki: ACM, p.96-103, 2008.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing of Environment**, v. 37, n. 1, p. 35–46, 1991.

ESRI. **ArcGIS Desktop: Release 10Redlands**. CA Environmental Systems Research Institute, 2011.

GHOSAL, V.; TIKMANI, P.; GUPTA, P. Face Classification Using Gabor Wavelets and Random Forest. Canadian Conference on Computer and Robot Vision Face. **Proc. in Canadian Conference on Computer and Robot Vision**, pp. 68-73, 2009.

MYERS, J. C. **Geostatistical error management. Qualifying uncertainty for enviromental sampling and mapping**. New York: Van Nostrand Reinhold, 1997. 571 p.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. Viena, Áustria, R Foundation for Statistical Computing, 2015. Disponível em: http://www.r-project.org/. Acesso em 24 mar 2014.

RICHARDS, J. A; JIA, X. **Remote Sensing Digital Image Analysis: An Introduction**. Berlim: Springer, 2006.

CAMPBELL, J. **Introduction to remote sensing**. 5. ed. New York: The Guilford Press, 1987.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo Agropecuário: Brasil, Grandes Regiões e Unidades da Federação**. Rio de Janeiro: IBGE, 2006.
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo Demográfico 2010**. Rio de Janeiro: IBGE, 2010.
RADAMBRASIL. **Levantamento exploratório de solos Folha Rio Doce**. Rio de Janeiro: IBGE, 1987.
RADAMBRASIL. **Folhas SF 23/24 Rio de Janeiro/ Vitória, geologia, geomorfologia, pedologia, vegetação e uso potencial da terra**. Rio de Janeiro: IBGE, 1983.
VALVERDE, O. Estudo regional da Zona da Mata de Minas Gerais. **Revista Brasileira de Geografia Física**, v. 20, n. 1, p. 3–32, 1958.
SCOLFORO, J. R.; CARVALHO, L. M. T. DE. **Mapeamento e inventário da flora nativa e dos reflorestamentos de Minas Gerais**. Lavras: UFLA, 2006.
UNIVERSIDADE FEDERAL DE VIÇOSA - UFV. **Levantamento de solos e aptidão agrícola da porção mineira da bacia do rio Doce**. Belo Horizonte: FEAM, 2010a.
UNIVERSIDADE FEDERAL DE VIÇOSA - UFV. **Levantamento de Solos e Aptidão Agrícola das Terras da Bacia do Rio Paraíba do Sul, Minas Gerais**. Belo Horizonte: FEAM, 2010b.