

Efeito das estratégias de treinamento na exatidão dos modelos preditivos em classificação de imagens

Leiliane Bozzi Zeferino¹
Ligia Faria Tavares de Souza¹
Elpídio Inácio Fernandes Filho¹
Teógenes Senna de Oliveira¹

¹ Universidade Federal de Viçosa - UFV/Departamento de Solos
Av. Peter Henry Rolfs, s/n, Campus Universitário - 36570-900 - Viçosa - MG, Brasil
{leiliane.zeferino, ligia.souza, elpidio, teo}@ufv.br

Abstract. This paper aims to evaluate the effect of training samples as points and polygons related to four different approaches by the Gradient Boosting Machine (GBM) and Random Forest (RF) classifiers on the accuracy of the supervised classification. The process of image classification used 11 predictive co-variables, six were spectral co-variable, and the remaining 5 were related to topography, climate, geology and pedology. The treatments were: 1 – separation of training and validation data without considering the origin polygon; 2 – separation of training and validation sample considering the origin polygons; 3 – random choice of a pixel from each polygon to represent this; and 4 – median and mode statistics calculated to each numeric and categorical co-variable, respectively. The results concluded that the treatments with the samples as points, treatments 3 and 4, presented the lowest average. Those with the sample as polygons diverged among itself, and treatment 1 presented the highest average. The validation with external data made the approaches became similar, where treatment 1 and 2, the same way as treatments 3 and 4, did not show any statistical difference, which can be concluded that this is necessary to the generation of more reliable predictive models. The effect of the sample collection in polygons to latter obtain a point, as approached by treatment 4, showed to be superior when compared to those that used the information as a whole polygon, even presenting lowest kappa index values.

Palavras-chave: remote sensing, land use and land cover map, Kappa index, sensoriamento remoto, mapa de uso e cobertura do solo, índice Kappa.

1. Introdução

O processo de classificação supervisionada de imagens é constituído basicamente por duas etapas. Na primeira etapa, amostras para treinamento do classificador são selecionadas para as diferentes classes de uso de interesse, permitindo ao classificador, já na segunda etapa, identificar padrões semelhantes aos apresentados pelas amostras coletadas e atribuir aos demais pixels da imagem, uma classe de acordo com as estatísticas específicas de cada algoritmo.

A validação é a etapa do processo de classificação que determina o grau de confiabilidade do modelo construído durante a fase de treinamento, com base no conjunto de dados inicial. Nesse sentido, destaca-se a importância da qualidade dos dados a serem usados, já que viés à classificação é inserido, principalmente, nessa etapa do processamento.

Nesse sentido, a seleção de estratégias de treinamento, deve considerar uma série de fatores, como o tamanho das amostras utilizadas no treinamento, o efeito da autocorrelação espacial, a variação ao longo da imagem, o tempo e os custos de mão-de-obra e a funcionalidade dos sistemas de processamento atuais (Campbell, 1996; Jensen, 1996). Entretanto, dificilmente consegue-se atender a todos os quesitos durante a coleta de amostras.

Dentre esses fatores Stehman & Wickham (2011) destacam o formato das amostras de referência, por influenciarem fortemente a avaliação da exatidão das classificações. Chen & Stow (2002), afirmam que o formato da amostra pode reduzir a capacidade dessas em representar uma classe, assim como favorecer a geração de resultados tendenciosos.

De posse dessas informações, o objetivo do trabalho é avaliar o efeito da coleta de amostras de treinamento em pontos e polígonos, associado a quatro abordagens de

treinamento dos algoritmos classificadores *Gradient Boosting Machine* (GBM) e *Random Forest* (RF) na exatidão da classificação supervisionada.

2. Metodologia de Trabalho

2.1. Descrição da Área de Estudo

A área de estudo é a bacia hidrográfica do Rio Lontra, localizada na região Norte do estado do Tocantins, inserida na Amazônia Legal e no sistema hidrográfico Tocantins-Araguaia. Nesta bacia está localizado um importante pólo econômico da região abrangendo os municípios de Aragominas, Araguaína, Araguaianã, Babaçulândia, Carmolândia, Piraquê, Wanderlândia e Xambioá (Figura 1).

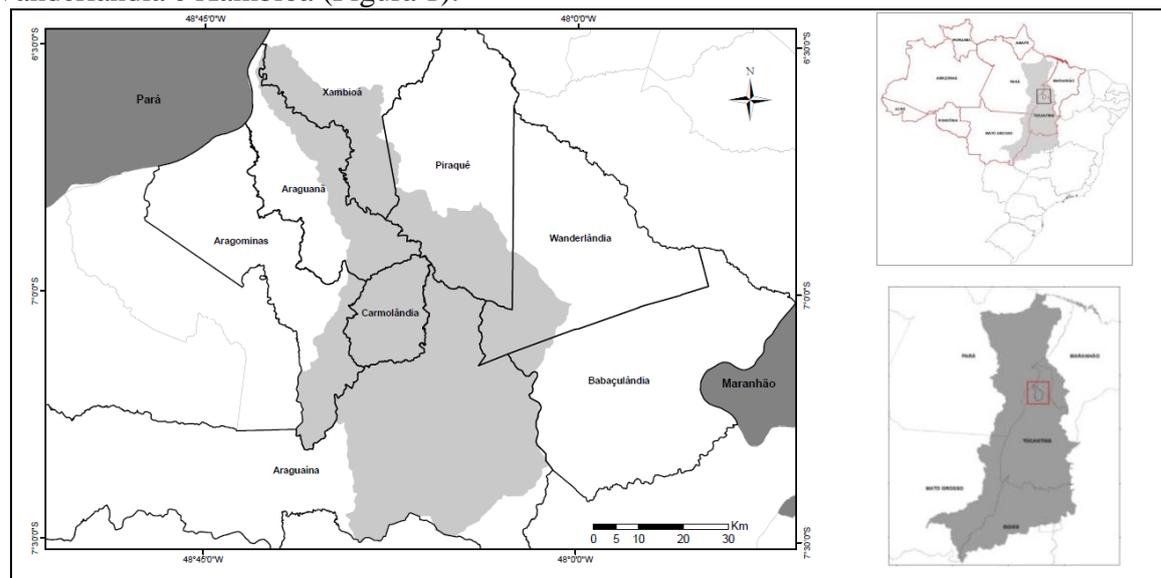


Figura 1. Mapa de localização da área de estudo no contexto na Amazônia Legal e sistema hidrográfico Tocantins - Araguaia.

Para a cobertura da área de estudo, foram selecionadas duas cenas do satélite Landsat 8 - OLI/TIRS, nas órbitas ponto 222/65 e 223/65. A fim de obter melhor distinção entre as possíveis classes de uso do solo foram utilizadas duas imagens, uma para o início do período seco e outra para o final do período seco. As datas de aquisição para cada uma dessas são apresentadas na tabela 1.

Tabela 1. Data de aquisição das imagens conforme a órbita/ponto para o período chuvoso e seco.

Órbita/Ponto	Data de Aquisição
<u>Início do Período Seco</u>	
223/65	15 de julho de 2015
222/65	24 de julho de 2015
<u>Final do Período Seco</u>	
223/65	18 de agosto de 2015
222/65	11 de agosto de 2015

2.2. Coleta de Amostras de Referência

Considerando as características espectrais de cada classe de uso e cobertura do solo, foram coletados, aleatoriamente, um conjunto de 606 polígonos de tamanho médio de 12 pixels da imagem Landsat 8 OLI, correspondendo a 7358 pixels/pontos amostrais. Na coleta, buscou-se contemplar as nove classes de uso e cobertura do solo, sendo elas: mata, cerrado típico, pastagem, queimada, área urbana, solo exposto, afloramento rochoso, floresta plantada de eucalipto e corpos d'água.

2.3. Co-Variáveis

O processo de classificação de imagens contou com 11 co-variáveis preditoras, das quais seis eram co-variáveis espectrais oriundas das imagens Landsat e as cinco restantes dizem respeito ao relevo, clima, geologia e pedologia (Tabela 2). Essas foram inicialmente selecionadas, de um conjunto de 99 co-variáveis, com base na avaliação da importância de cada para geração do modelo de previsão de uso e cobertura do solo, usando para tanto o algoritmo *Random Forest*.

Tabela 2. Descrição das co-variáveis utilizadas no processo de classificação de imagem

Identificação	Nome da Co-Variável	Tipo da Co-Variável
1	SAVI ^{1/} imagem início período seco	
2	Diferença SAVI: SAVI início período seco – SAVI final período seco	
3	Banda 2 (azul): imagem início período seco	
4	Banda 5 (infravermelho próximo): imagem início período seco	Espectral
5	Banda 4 (vermelho): imagem final período seco	
6	Banda 5 (infravermelho próximo): imagem final período seco	
7	Declividade	Relevo
8	Mapa Geológico	Geológica
9	Mapa de Solos	Pedológica
10	Precipitação do mês de Abril	Climática
11	Precipitação do mês mais chuvoso	Climática

^{1/} SAVI : Índice de Vegetação Ajustado ao Solo

2.4. Métodos para seleção de amostras de treinamento e validação

O processo consistiu de 10 ciclos consecutivos de reamostragem para a geração aleatória de conjuntos de treinamento e validação. Para tanto, o conjunto amostral total constituído por 606 polígonos (correspondendo a 7358 pontos) foi particionado a cada ciclo, em treinamento e validação na proporção 0,75: 0,25, respectivamente. Os conjuntos de treinamento foram utilizados como dados de entrada para geração dos modelos preditivos em cada tratamento, enquanto os conjuntos de validação foram utilizados como validadores externos aos modelos para aferir a exatidão dos mesmos (Figura 2).

A validação interna adotou a metodologia da validação cruzada com *K-folds* (subconjuntos). Nesse método, o conjunto de treinamento oriundo em cada ciclo de partição, como citado anteriormente, foi dividido em *K* subconjuntos. Destes *K* subconjuntos, um foi retido para ser usado na validação interna do modelo e os *K-1* subconjuntos restantes foram utilizados no treinamento. Assim, o processo de validação cruzada foi repetido *K* vezes, de modo que cada um dos *K* subconjuntos fossem utilizados uma vez como dado de teste para validação do modelo. Para tanto, adotou-se para esse trabalho, *K* = 5 subconjuntos e, portanto, quatro subconjuntos foram utilizados para treinamento e um reservado para validação interna.

Os tratamentos consistiram da associação de duas formas das amostras, polígono e pontos (pixel), a quatro esquemas diferentes de separação entre amostras de treinamento e validação do conjunto de dados inicial para o processo de validação interna (Figura 3).

No primeiro tratamento foram separados dados de treinamento e validação sem considerar o polígono de origem. Através dessa abordagem, cada um dos 12 pixels que constituem um polígono é tratado como amostra única e, portanto, o conjunto de treinamento e validação podem apresentar pixels oriundos de uma mesma amostra de origem, ou seja, de um mesmo polígono (Figura 3A).

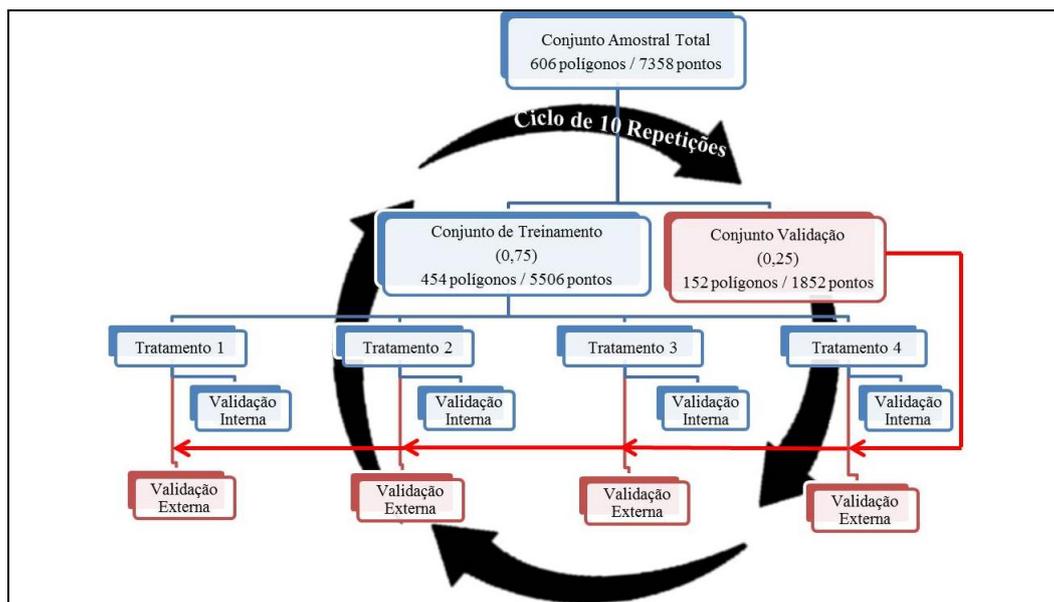


Figura 2. Fluxograma do processamento dos tratamentos abordados.

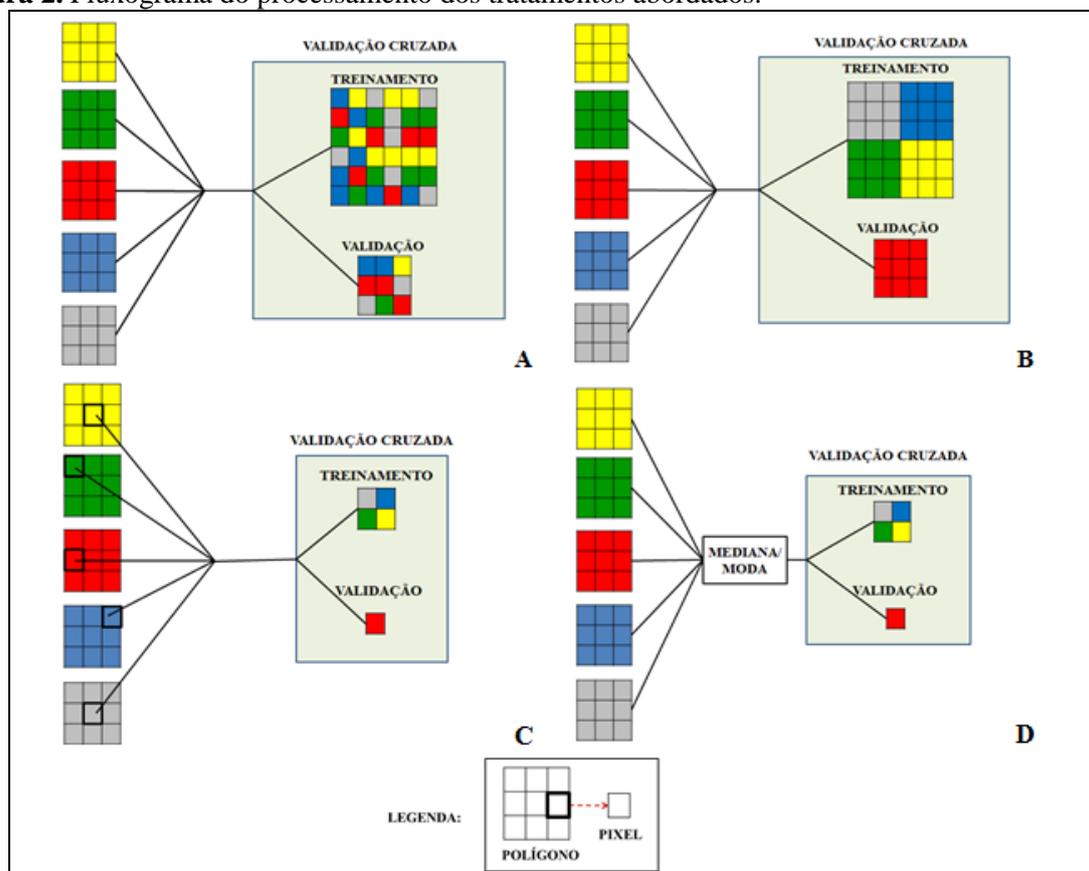


Figura 3. Esquematização dos tratamentos abordados. A- Tratamento 1; B- Tratamento 2; C- Tratamento 3; e D- Tratamento 4.

O segundo tratamento consistiu da separação de amostra de treinamento e validação com base nos polígonos, onde todos os 12 pixels amostrais correspondentes a um polígono são alocados no conjunto de dados de treinamento ou validação. Assim, nesse tratamento são gerados conjuntos de treinamento e validação mutuamente exclusivos em termos de polígonos amostrados (Figura 3B).

No terceiro tratamento foi realizada a escolha aleatória de um pixel para cada polígono, passando esse, a representar o polígono (Figura 3C). No quarto e último tratamento, foram

calculadas as estatísticas mediana e moda para cada uma das co-variáveis, numéricas e categóricas, respectivamente. Dessa maneira, assim como na abordagem anterior, um pixel que apresenta o valor mediano ou modal, representará o polígono (Figura 3D). Na tabela 3 é representado o tamanho dos conjuntos de treinamento e validação para cada tratamento abordado.

Tabela 3. Tamanho dos conjuntos amostrais conforme os tratamentos abordados

Tratamento	Conjunto Amostral (pixels)	Amostra (pixels)	Validação Externa (pixels)
1	7358	5506	1852
2	7358	5506	1852
3	7358	454	1852
4	7358	454	1852

2.5. Classificadores utilizados

Os algoritmos classificadores utilizados foram *Generalized Boosted Models* (GBM) e *Random Forest* (RF), ambos pertencem ao grupo de classificadores que fazem uso do método de árvores de decisão.

O classificador GBM funciona por meio da realização de sucessivas árvores, sendo que a cada árvore gerada, pesos são distribuídos a cada classe conforme ao erro obtido durante a predição das mesmas, priorizando àquela que obteve maior erro. Após consecutivos ciclos de treinamento, o classificador acaba por promover “*overfitting*” no treinamento, ou seja, superajuste. Quando atingido esse estágio, o classificador seleciona a árvore com modelo de melhor ajuste e equilíbrio na distribuição de pesos às classes.

Já o classificador RF, atua através da geração de várias árvores simultâneas. Para determinar a classe de um objeto, o método combina o resultado das árvores através da concessão de votos. Assim, a classificação final é dada pela classe que recebeu o maior número de votos entre todas as árvores geradas pelo algoritmo (Diniz et al., 2013).

3. Resultados e Discussão

O resultado da análise de variância revelou a existência de diferença significativa para os fatores - tratamento, classificador e tipo de validação - e para a interação entre os fatores tratamento e validação. Para as fontes de variação significativas, procedeu-se análise de comparação múltipla entre médias através do Teste Tukey ($\alpha = 0,05$), para as médias dos fatores e Teste Z ($\alpha = 0,05$) para as médias da interação tratamento e validação.

Para o fator tratamento conforme apresentado no Gráfico 1, nota-se que os tratamentos que adotam a forma da amostra em ponto, tratamento 3 e 4, apresentam as menores médias, mas semelhantes entre si. Já aqueles que abordam a amostra na forma de polígonos, diferiram entre si, tendo o tratamento 1 a maior média. Ressalta-se que estão, neste caso, sendo avaliados os efeitos isolados dos níveis dos fatores do tratamento, desconsiderando os efeitos dos classificadores e do tipo de validação.

A superioridade do tratamento 1 é consequência da falta de controle no processo de separação dos conjuntos de treinamento e validação interna. Assim, cada pixel de um polígono foi considerado uma amostra individualizada, permitindo que pixels de um mesmo polígono fossem agrupados parte no conjunto de treinamento e parte no conjunto de validação. Dessa forma, os resultados obtidos nessa abordagem revelam-se tendenciosos conforme relatado por Hammond & Verbyla (1996), Franklin (1994) e Foody & Mathur (2006). Segundo esses autores, a obtenção de resultados confiáveis com ausência de viés depende de uma validação com amostras diferentes das utilizadas para o treinamento do classificador, o que não pode ser garantido nesse primeiro tratamento, como demonstrado no esquema da Figura 3.

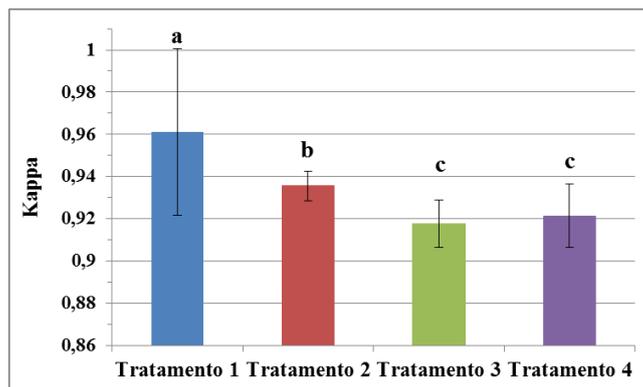


Gráfico 1. Comparação das médias do índice kappa para os níveis do fator tratamentos. As médias seguidas pela mesma letra, não diferem entre si pelo Teste Tukey ao nível de 0,05 de significância.

Tal comportamento é confirmado, através da diferença significativa entre os valores de índice kappa para validação interna e externa. A presença de amostras na validação externa, diferente daquelas utilizadas na etapa de treinamento, permitiu remover o viés relatado anteriormente, o que por sua vez, resultou em menores valores de índice kappa comparados aos apresentados pelo conjunto de dados utilizados na validação interna (Gráfico 2). No Gráfico 2 nota-se que a validação interna, independente dos demais fatores, apresentou médias significativamente maiores, possivelmente em razão dos valores do tratamento 1.

A interação também significativa, entre os fatores tratamento e validação auxiliam na comprovação do efeito da presença de amostras de mesma origem nos conjuntos de treinamento e validação. Esses aspectos corroboram para a importância não só da qualidade dos dados para treinamento do classificador (Chen & Stow, 2002), mas também para o modo como as técnicas de treinamento e validação são conduzidas, o que poderão se tornar fontes de viés para a classificação.

A mudança na classificação após a utilização da validação externa pode ser notada no Gráfico 2. A validação com dados externos tornou as abordagens semelhantes entre si, de maneira que os tratamentos 1 e 2, bem como os tratamentos 3 e 4, não diferiram estatisticamente.

Por outro lado, estudos (Foody & Mathur, 2004; Pal & Mather, 2003; Zhuang et al., 1994; Chen & Stow, 2002) tem se dedicado a avaliar os efeitos do tamanho do conjunto de treinamento na exatidão das classificações. Para esses autores, os resultados dos índices de exatidão estão diretamente relacionados ao tamanho do conjunto amostral.

Nesse sentido é possível compreender a similaridade dos resultados para aqueles tratamentos que assume as amostras na forma de ponto, tratamento 3 e 4 (Gráfico 1). Conforme apresentado na tabela 4, tais tratamentos possuem o conjunto de treinamento 12 vezes menor que aqueles que consideravam a amostras na forma de polígono, tratamento 1 e 2.

Para Stehman & Czaplewski (1998), não há ainda consenso sobre o melhor formato da amostra, sendo que a escolha de um formato depende do objetivo da pesquisa, das características da paisagem, características do processo de mapeamento, bem como as restrições práticas de cada forma. Esses autores ainda acrescentam que a escolha deve ser ponderada entre os benefícios e custos, principalmente operacionais, que estão associados a cada tipo de amostragem.

O efeito da coleta de amostras em polígonos, com posterior processamento para obtenção de um ponto, como abordado no tratamento 4, revela-se nitidamente superior àquelas abordagens que utilizam as informações de um polígono por completo. Esse efeito pode ser justificado devido a melhor representação daquelas classes com maior variabilidade entre amostras e dentro da amostra.

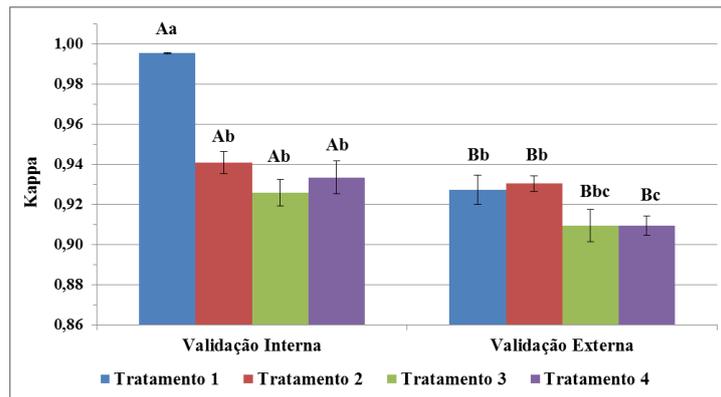


Gráfico 2. Valores médios do índice kappa para a interação do fator tratamento e do fator validação. As médias seguidas pela mesma letra em maiúsculo não diferem entre si para os níveis do fator validação. As médias seguidas pela mesma letra em minúsculo não diferem entre si para a interação do fator tratamento e fator validação. Teste Z ($\alpha = 0,05$).

Assim, utilizando o valor da estatística mediana e moda, assume-se o comportamento de um conjunto de n -pixels, conforme o tamanho do polígono, em um ponto. Com isso, é possível reduzir o esforço computacional para geração de modelos preditos durante o processo de obtenção de mapas de uso e cobertura do solo (Chen & Stow, 2002), e ao mesmo tempo manter resultados de índice de exatidão elevados, com menor variação entre as repetições.

Nesse contexto, Hammond & Verbyla (1996) relatam a importância da coleta de amostras em áreas heterogêneas, uma vez que restringir a avaliação a áreas homogêneas pode ser a causa de resultados de índices de exatidão otimistas. Campbell (2003) e Mather (2004) acrescentam que as amostras devem contemplar a variabilidade natural da classe a ser mapeada, a fim de ser capaz de descrever a população geral da classe que representa (Foody & Mathur (2004).

Além desses aspectos, o tratamento 4 ao adotar medidas estatísticas tais como a mediana e moda, remove o efeito da auto correlação espacial existente na coleta de amostras contíguas ou próximas entre si (Campbell, 1981; Labovitz & Masuoka, 1984; Gong & Howarth, 1990), como nas abordagens realizadas nos tratamentos 1 e 2, que utilizam todos os pixels de polígono na obtenção de informações e reconhecimento de padrões. Portanto, por se tratarem de pixels contíguas, há nesses casos a presença marcante do efeito de correlação espacial entre os pixels. As consequências desse viés são a geração de modelos não representativos, bem como a obtenção de índice kappa superestimado.

Dessa forma, torna-se nítido que adotar medidas estatísticas, como mediana e moda, as quais representam a tendência central dos dados - mediana, bem como o valor predominante - moda, permitem a geração de resultados mais fidedignos às características do ambiente a ser mapeado, uma vez que consideram a variabilidade espacial das características das classes de interesse. Comportamento oposto a esse pode ser observado no tratamento 3, o qual por adotar apenas as informações de um pixel para representar uma classe, perde a capacidade de reproduzir a possível variabilidade natural de uma classe.

4. Conclusões

Os resultados apresentados permitem concluir que o índice kappa é influenciado pelo formato da amostra de referência e pelas técnicas como são conduzidas as etapas de treinamento e validação. Apesar de ter sido estatisticamente diferentes entre si, sendo as médias para o RF maiores, os classificadores não apresentaram relação com os demais fatores avaliados.

Em relação os formatos de amostragem, destaca-se a abordagem em pontos do tratamento 4, mesmo apresentando as menores exatidões. Nesse tratamento, há ganho operacional ao se adotar as medidas estatísticas mediana e moda para a obtenção de ponto, reduzindo o conjunto de dados em 12 vezes. Outra vantagem, é que por esse tratamento torna-se possível considerar a heterogeneidade natural dos objetos, bem como a remoção da correlação espacial existente entre as amostras, o que não é possível nos tratamentos anteriores.

Por fim, a utilização de validação com dados externos mostra-se ser uma ferramenta necessária para a geração de modelos mais confiáveis no processo de classificação de imagens.

Referências Bibliográficas

- CAMPBELL, J.B. Introduction to Remote Sensing, The Guilford Press, New York, N.Y., 622 p., 1996.
- CAMPBELL, J.B. Spatial correlation effects upon accuracy of supervised classification of land cover, Photogrammetric Engineering and Remote Sensing, 47(3):355-363, 1981.
- CAMPBELL, J. B. Introduction to remote sensing. The Guilford Press, London, 3rd edition, 621p., 2003.
- CHEN, D., STOW, D. The Effect of Training Strategies on Supervised Classification at Different Spatial Resolutions. Photogramm. Eng. Remote Sens. 68, 1155–1161, 2002.
- DINIZ, F. A. ; MENDES NETO, F. M.; FONTES JÚNIOR, F. C. L., LAYSA, M. O. F.; RedFace: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces: comparação com diferentes classificadores. Revista Brasileira de Computação Aplicada (ISSN 2176-6649), Passo Fundo, v. 5, n. 1, p. 42-54, 2013.
- FOODY, G.M., MATHUR, A. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103, 179–189, 2006.
- FOODY, G.M., MATHUR, A. Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification. Remote Sens. Environ. 93, 107–117, 2004.
- FRANKLIN, S. Discrimination of subalpine forest species and canopy density using digital CASI, SPOT PLA, and Landsat TM data. Photogrammetric Engineering and Remote Sensing, 60, 1233-1241, 1994.
- GONG, P., & HOWARTH, P.J. An assessment of some factors influencing multispectral land cover classification, Photogrammetric Engineering and Remote Sensing, 56(5):597-603, 1990.
- HAMMOND, T.O., VERBYLA, D.L. Optimistic bias in classification accuracy assessment. Int. J. Remote Sens. 17, 1261–1266, 1996.
- JENSEN, J.R. Introductory Digital Image Processing: A Remote Sensing Perspective, Prentice Hall Series in Geographic Information Science, Prentice Hall, Upper Saddle River, New Jersey, 316 p., 1996.
- LABOVITZ, M.L., & MASUOKA, E.J. The influence of autocorrelation in signature extraction-An example from a geobotanical investigation of Cotter Basin, Montana, International Journal of Remote Sensing, 5(2):315-332, 1984.
- MATHER, P. M. Computer processing of remotely sensed images. Chichester: Wiley, 3rd edition 442 p., 2004.
- PAL, M., & MATHER, P. M. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, 554– 565, 2003.
- STEHMAN, S. V, CZAPLEWSKI, R.L. Design and Analysis for Thematic Map Accuracy Assessment - an application of satellite imagery. Remote Sens. Environ. 64, 331–344, 1998.
- STEHMAN, S.V., WICKHAM, J.D. Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. Remote Sens. Environ. 115, 3044-3055, 2011.
- ZHUANG, X., ENGEL, B. A., LOZANO-GARCIA, D. F., FERNANDEZ, R. N. & JOHANNSEN, C. J. Optimisation of training data required for neuro-classification. International Journal of Remote Sensing, 15, 3271–3277, 1994.