Comparação no mapeamento da cultura de milho safrinha utilizando *Machine Learning* em imagens Landsat-8

Luiz Almeida [1]
Jerry Adriani Johann [1]
Jonathan Richetti [1]
Rafaela Fernandes Nicolau [1]
Amanda Bordin Richetti [1]

[1] Universidade Estadual do Oeste do Paraná - UNIOESTE
Rua Universitária, 20169 – Cascavel – PR, Brasil
{almeidalz, jerry.johann, j_richetti, rafa.nicolau}@hotmail.com;
amandabitaliano@gmail.com

Abstract. The objective of this study was to compare the mapping of winter corn, using Machine Learning in Landsat-8 images in 2016 crop. For the images processing the software R 3.3.1 and ArcMap 10.0 were used. From a false-color RGB-564 composition of the Landsat-8 images 5 classes of soil use and cover (urban area, water bodies, forest, winter corn and exposed soil) were polygonised. These sampled areas served as training data for the models. The Random Forest and the Gamboost classification methods were applied. To perform the accuracy of each mask random points were generated for each classification and a being point-to-point verification was performed. For the Gamboost method the value of the adjustment parameter that allowed the best result was 150 iterations (Mstop). While Random Forest presented the best classification result when the number of predictors sampled in each node (Mtry) was equal to 2. The winter corn area identified in each model was about 75,290.58 ha for GB and 57,220.29 ha for RF, with Global Accuracy of 87.75% and 79.0%, respectively. In spite of the differences between the classifiers used, both methods are effective in mapping the studied culture. Moreover, both methods presented great agility to classify and to obtain area, aiding in the ergonomics of the processes.

Palavras-chave: Gamboost, Random Forest, satellite image processing, sensoriamento remoto.

1. Introdução

O milho safrinha está entre as culturas de maior interesse e valor econômico no Brasil, destacando-se como um dos principais produtos da agricultura nacional na safra de inverno (IBGE, 2002). O estado do Paraná destaca-se no cultivo da cultura, representando uma produção de 20% em 2015 e 26% em 2016 em relação a produção nacional (IBGE, 2016).

A busca de metodologias que quantifique de maneira rápida, eficaz e de baixo custo a estimativa de safra auxilia no monitoramento das culturas agrícolas. Neste sentido, o sensoriamento remoto mostra-se como uma ferramenta útil na coleta de dados e o uso de imagens satélites permitem a identificação e quantificação de informações, de forma ágil auxiliando no levantamento de dados agrícolas.

Dentre as metodologias de obtenção de dados de áreas de imagens de satélite, as ações de classificação possibilitam a criação de mapas digitais temáticos de áreas. Segundo Crosta (2002), as metodologias de classificação supervisionada, com o uso de algoritmos classificadores, necessitam de prévio conhecimento de regiões das imagens, de modo a servir de amostra para comparação com as demais regiões; essas metodologias geralmente apresentam melhor desempenho no processo de classificação.

Portanto, o objetivo deste trabalho foi realizar a comparação nos mapeamentos para a cultura de milho safrinha, safra 2016, utilizando classificadores do tipo *Machine Learnig*, mais especificamente, gerar o mapa de classificação pelo método *Gamboost* e compará-lo com o mapa de classificação gerado pelo método *Random Forest*, utilizando imagens do satélite Landsat-8.

2. Metodologia de Trabalho

A área em estudo, compreende o *tile* 223/77 (órbita/ponto) do satélite Landsat-8, compreendido entre os paralelos 23°30'S e 25°35'S e os meridianos 51°55'W e 54°10'W (Figura 1). A data do *tile* (imagem) foi escolhida com base nas datas de plantio de milho safrinha para o estado do Paraná, sendo 30 de abril de 2016, essa data indica que 100 % da cultura já foi semeada na região de estudo (SEAB/DERAL, 2016). Para o processamento das imagens foram utilizados os softwares R 3.3.1 (R CORE TEAM, 2016) e ArcMap 10.0.

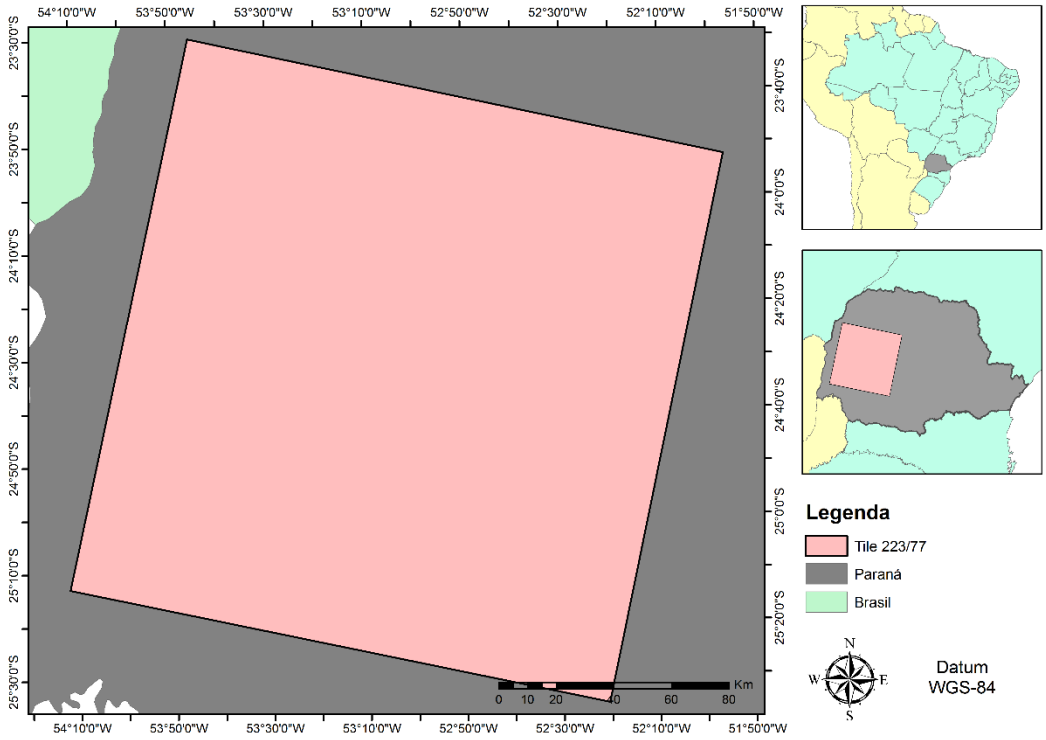

Figura 1. Área de estudo, *tile* 223/77 (órbita/ponto) Landsat-8, sobre o estado do Paraná.

Previamente foi gerada uma composição falsa-cor RGB-564 e definidas 5 classes de uso e cobertura do solo (área urbana, corpos hídricos, mata, milho safrinha e solo exposto). Para cada classe, foram selecionadas regiões amostrais, bem definidas, e virtualmente poligonizadas, sendo exportadas, em arquivo único, para formato ESRI *shapefile* (shape) no ArcMap 10.0. Esse resultante passou a ser os dados de treinamento (*train data*), com atribuições de valores binários para as classes.

Com o software R 3.3.1 (R CORE TEAM, 2016), na sequência de imagens das bandas de 2 a 7 do *tile* foram aplicados os métodos de classificação *Random Forest* e *Gamboost*, com o auxílio dos os pacotes *caret* (Kuhn et al, 2016); *ggplot2* (Wickham et al., 2016); *mboost* (Hothorn et al., 2016); *raster* (Hijmans, 2016) e *rgdal* (Bivant et al., 2016). Não houve pré-processamento das imagens e cada método foi executado independentemente um do outro, sendo gerados mapas de classificação, de cada classificador para a cultura de milho safrinha.

O método de *machine learning Random Forest*, consiste em um conjunto de classificadores (*bagging*), onde cada classificador (árvore) é treinado e seus resultados individuais sofrem um processo de seleção por "votação". Dessa votação é eleita uma classe "mais popular", sendo aplicada novamente nas árvores, realizando esse procedimento k' vezes (iterações) quantas forem de necessidade ou solicitação (*Mtry*) (Breiman, 2001). Com esse método gerou-se uma máscara para a safra 2016 de milho (máscara RF).

Já o método *Gamboost* é um algoritmo de classificação *machine learning*, do tipo *boosting* aditivo de modelos generalizados (*GAM*), podendo-se resumir que, as generalidades são estimadas por um gradiente "componente-sábio" e as adições pelas cotas iteráveis que reforçam o modelo. Esse reforço ou número de iterações de aumento inicial (*Mstop*), é interpretado como um aditivo da função de predição (Hofner, 2014). Com esse método gerou-se outra máscara para a safra 2016 de milho (máscara GB).

Além das métricas disponíveis para cada classificador, acurácia e Kappa, para o *train data*, houve a análise de acurácia das máscaras. Para realização de acurácia das máscaras, foram gerados pontos aleatórios em cada classificação. Realizou-se a verificação ponto a ponto, observando se o ponto pertencia ou não a cultura em estudo. Com base nas verificações ponto a ponto, elaborou-se a matriz de confusão, possibilitando os cálculos de Exatidão Global (EG) (Equação 1) e o Índice Kappa (Equação 2).

$$EG\ (\%) = \frac{A}{m} * 100 \quad (1)$$

$$k = \frac{N * \sum_{i=1}^{r} Xii - \sum_{i=1}^{r} (Xi+ * X+i)}{N^2 - \sum_{i=1}^{r} (Xi+ * X+i)} \quad (2)$$

Em que n é número de observações (pixels amostrais); A é acerto geral (pixels amostrais classificados corretamente); m é número de pixels amostrais; r é número de linhas da matriz de erro; x_{ij} é observações na linha i e coluna i; x_i é total marginal da linha i; x_j é total marginal da coluna j.

3. Resultados e Discussão

A Figura 2 apresenta as classificações para milho safrinha, safra 2016, *tile* 223/77, em imagens Landsat-8: A) máscara gerada pelo método *Gamboost*; B) máscara resultado do método *Random Forest*.

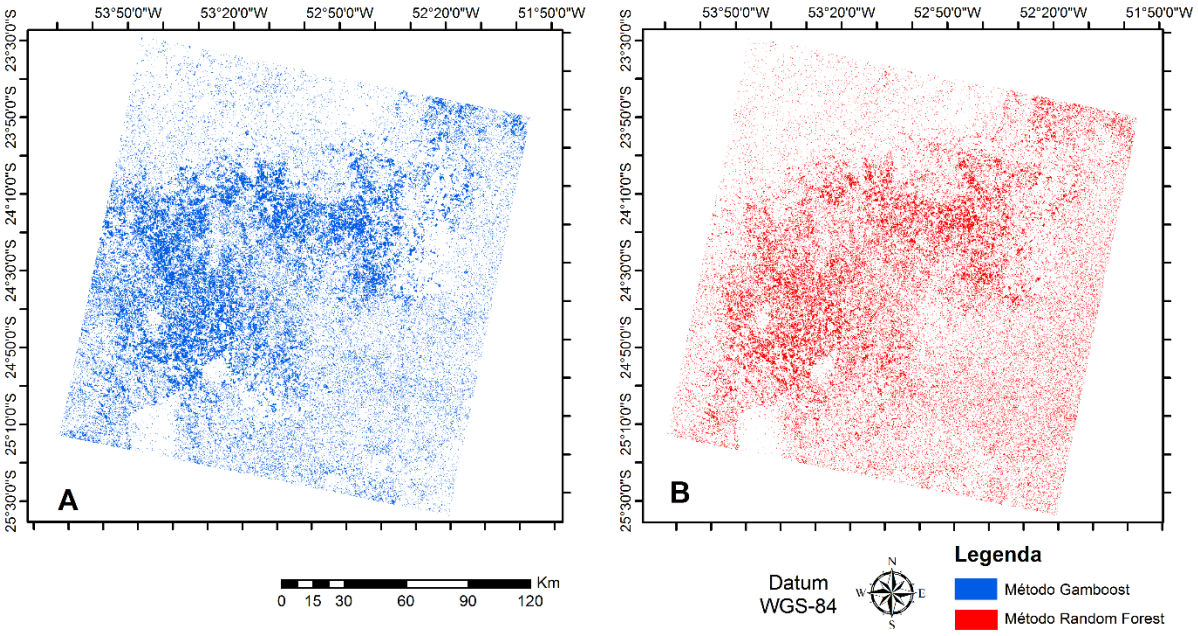

Figura 2. Resultados dos métodos de classificação do *tile* 223/77 para milho safrinha 2016.

Visualmente, pode-se verificar maior densidade de área classificada pelo método *Gamboost* (Figura 2 - A), em relação ao outro método. Realizando a extração de áreas das máscaras, obteve-se um total de 75.290,58 hectares classificados com o *Gamboost*, enquanto *Random Forest* identificou 57.220,29 hectares; evidenciando a afirmação anterior.

Observa-se, nos parâmetros de ajuste e métricas avaliadoras de cada classificador, para os dados de treinamento, bem como estatísticas para as máscaras (Tabela 1), que para geração do mapa de classificação com o método *Gamboost*, o valor do parâmetro de ajuste que possibilitou melhor resultado foi de 150 iterações (*Mstop*). Com acurácia de 0,7251 e Índice Kappa igual a 0,4231, para o *train data*. Contudo, o Índice Kappa resultante da avaliação de acurácia da máscara GB foi 0,7550, com Exatidão Global de 87,75 %.

Já o modelo *Random Forest*, apresentou o melhor resultado de classificação quando o número de preditores amostrados em cada nó (*Mtry*) foi igual a 2. Isso possibilitou Índice Kappa e acurácia, para o *train data*, de 0,8228 e 0,9154 respectivamente. Enquanto, a avaliação da máscara RF apresentou Índice Kappa de 0,58 com Exatidão Global de 79,00 %.

Tabela 1. Estatísticas dos classificadores para os dados de treinamento e máscaras.

			Dados de treinamento		Classificação	
Método	**Parâmetro**		**Acurácia**	**Kappa**	**Exatidão Global**	**Kappa**
Gamboost	**Mstop**	50	0,7141	0,3959	-	-
		100	0,7181	0,4066	-	-
		150	0,7251	0,4231	87,75%	0,7550
Random Forest	**Mtry**	2	0,9154	0,8228	79,00%	0,5800
		4	0,9118	0,8156	-	-
		6	0,9055	0,8030	-	-

Realizando a comparação com trabalhos que utilizaram classificadores *machine learning* dos tipos *bagging* e *boosting*, Chan et al. (2008) encontrou valores de acurácia de 0,6880 e 0,6950 para os métodos *Random Forest* e *Adaboost*, respectivamente, onde avaliou a classificação de cada modelo no mapeamento de regiões ecologicamente homogêneas em imagens de satélite (*ecotope*). Já Zhong et al. (2016), no mapeamento de milho e soja com base na fenologia, obteve uma acurácia de 87,2 % e Índice Kappa de 0,8040. Enquanto Johann (2012), utilizando imagens de máximos e mínimos de EVI , atingiu Exatidão Global de 94,72 % e Índice Kappa de,8945, no mapeamento de áreas com as culturas de verão no estado do Paraná, dentre elas o milho.

4. Conclusões

O método *Gamboost* apresentou melhor desempenho para o mapeamento da cultura de milho safrinha em imagens Landsat-8, em relação ao método *Random Forest*.

Apesar das diferenças de resultados entre os classificadores utilizados neste trabalho, os dois métodos mostram-se eficazes no mapeamento da cultura estudada e de grande agilidade para obtenção de dados de área, otimizando o tempo do analista.

Agradecimentos

Ao Laboratório de Estatística Aplicada (LEA) e ao Laboratório de Topografia e Geoprocessamento (GeoLab), ambos da UNIOESTE-Campus Cascavel, pela infraestrutura disponibilizada para a realização da pesquisa.

Referências Bibliográficas

Benjamin Hofner, Andreas Mayr, Nikolay Robinzonov and Mattthias Schmid (2014). Model-based Boosting in R - A Hands-on Tutorial Using the R Package mboost. **Computational Statistics**, v.29, p. 3-35, 2014.

Breiman L. Random Forest. **Machine Learning**, v.45, p. 5-32, 2001.

Chan J. C. –W.; Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. **Remote Sensinf of Environment**, v.112, p 2999-3011, 2008.

Crósta, A.P. **Processamento digital de imagens de sensoriamento remoto**. Campinas: IG/UNICAMP, 2002. 170 p.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Relatórios metodológicos.** Rio de Janeiro: Pesquisas Agropecuárias - Departamento de Agropecuária, 2002. 92 p.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Levantamento Sistemático da Produção Agrícola.** (2014). Disponível em: <http://www.sidra.ibge.gov.br/>. Acesso em: 28.out. 2016.

Johann, J. A. **Calibração de dados agrometeorológicos e estimativa de área e produtividade de culturas agrícolas de verão no estado do Paraná**, Tese de Doutorado. Campinas: Universidade Estadual de Campinas - UNICAMP, 2012.

Handley Wickham and Wiston Chang (2016). ggplot2: An Implementation of the Grammar of Graphics. R package version 2.1-0. Disponível em: <https://CRAN.R-project.org/package=ggplot2/> Acesso em: 3.out.2016.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan (2016). caret: Classification and Regression Training. R package version 6.0-71. Disponível em: <https://CRAN.R-project.org/package=caret/>. Acesso em: 3.out.2016.

Oger Bivand, Tim Keitt and Barry Rowlingson (2016). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.1-10. Disponível em: <https://CRAN.R-project.org/package=rgdal/>. Acesso em: 3.out.2016.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em: 23.set.2016.

Robert J. Hijmans (2016). raster: Geographic Data Analysis and Modeling. R package version 2.5-8. Disponível em: <https://CRAN.R-project.org/package=raster/>. Acesso em: 3.out.2016.

Secretaria da Agricultura e do Abastecimento (SEAB) – Departamento de Economia Rural (DERAL). Disponível em: <https://www.agricultura.pr.gov.br/>. Acesso em: 30.set.2016.

Torsten Hothorn, Peter Buhlmann, Thomas Kneib, Mattthias Schmid and Benjamin Hofner (2016). mboost: Model-based Boosting. R package version 2.6-0. Disponível em: <https://CRAN.R-project.org/package=mboost/> Acesso em: 3.out.2016.

Zhong, L; Hu, l; Yu, L; Gong, P, Biging, G. S. Automated mapping of soybean and corn using phenology. ISPRS Journal of Photogrammetry and Remote Sensing, v.119 p. 151-164, 2016.