Técnicas de mineração de dados aplicadas a imagens MODIS para mapeamento de culturas de verão no estado do Paraná.

Weverton Rodrigo Verica[1]
Clóvis Cechim Júnior[1]
Jonathan Richetti[1]
Laíza Cavalcante de Albuquerque Silva [1]
Willyan Ronaldo Becker [1]
Alex Paludo [1]
Jerry Adriani Johann[1]

[1] Universidade Estadual do Oeste do Paraná - UNIOESTE
R. Universitária, 2069 - Caixa Postal 711
85819-110 - Cascavel - PR, Brasil
{wevertonverica, juniorcechim, j_richetti, laiza.cavalcante, willyanbecker, paludo.alex
jerry.johann }@hotmail.com

Abstract. The objective of this work was to develop a methodology for mapping cultivated areas with summer crops (soybean and corn) in the state of Paraná using MODIS sensor images for the crop year of 2013/2014. For classification the Random Forest algorithm was used. It is hard to differ soybean from corn with MODIS. Due to the heterogeneous and high spectral-temporal dynamics of corn and soybean, including proximity or distinction in the sowing and initial development in mesoregions of the state the Random Forest algorithm was applied in order to present a clear differentiation between crops. For the evaluation of the spatial accuracy of the mapping, the Landsat8 / OLI satellite images were used. These images served as reference to generate the error matrix. The proposed method obtained a kappa index of 0.9678, which is considered excellent, and a global of 98.61%, which also represents an excellent index. However, the results obtained from the mapping underestimated the area destined to the culture of Corn in about 50% and overestimated the soybean area by about 24%. Besides that, the methodology was successful in mapping soybean and corn crops in the state of Paraná for the crop year 2013/2014, since the method is fast and inexpensive. Thus, the results indicate that the method is efficient for mapping the summer crops. Nevertheless, it is necessary to make some improvements to minimize the difference between the mapping result and the official data.

Palavras-chave: image classification, remote sensing, Random Forest, classificação de imagens, sensoriamento remoto, Random Forest

1. Introdução

Diante do grande crescimento da população mundial e o aumento da demanda por alimento, existe uma necessidade em expandir a produção mundial. Segundo Johann (2012) o Brasil é um dos países que apresentam as melhores condições para essa expansão.

No cenário brasileiro, o Paraná ocupa um lugar de destaque no meio agrícola sendo o segundo maior produtor de grãos no país. O estado do Paraná ocupa uma área de 199.880 km² e no ano de 2015 produziu 17,145 milhões de tonelada de soja e 15,973 milhões de toneladas de milho (IBGE, 2016).

Com isso nota-se a importância do estado do Paraná para a agricultura brasileira, principalmente no que se refere as culturas de milho e soja. Entretanto para que possa haver melhoria na produção agrícola do estado é essencial que haja planejamento e investimento nesse setor, porém para realizar essas tarefas com qualidade é necessário conhecer as áreas cultivadas na região de interesse.

Esse conhecimento pode ser adquirido por meio de interpretação visual, porém, para áreas extensas esse processo demanda tempo e custo, um método mais ágil para realizar o mapeamento de culturas é a partir de sensoriamento remoto orbital que utiliza imagens de

satélite para áreas agrícolas, conforme o trabalho de Johann et al., 2012 que tinha como objetivo estimar e mapear as áreas com as culturas de soja e milho, no estado do Paraná, utilizando de imagens de satélite, da mesma forma Santos et al., 2014, buscava mapear área cultivada de soja na região norte do Rio Grande do Sul através de imagens de satélite, também pode-se citar o trabalho de Brown et al., 2013, que classifica os dados plurianuais de uso da terra agrícola de Mato Grosso usando os índice de vegetação obtidos pelo sensor MODIS, entre outros trabalhos.

Para conseguir identificar as áreas cultivadas com soja e milho buscou-se técnicas de mineração de dados nas imagens obtidas dos sensores foram aplicadas. A mineração de dados é uma fase do processo de Descoberta de Conhecimento em Bases de Dados, do inglês "*Knowledge Discovery in Databases – KDD*", que também possui as fases de coleta de dados, pré-processamento, formatação e avaliação.

O KDD é um processo não trivial de descoberta de padrões válidos, novos, úteis e acessíveis, cujo principal vantagem do processo de descoberta é que não são necessárias hipóteses, sendo que o conhecimento é extraído dos dados sem conhecimento prévio. As principais tarefas de mineração de dados estão relacionadas a classificação, associação e agrupamento de padrões (Fayyad et al., 1996).

De acordo com Tan et al. (2009) a associação é utilizada para descobrir padrões que descrevam características altamente associadas em bancos de dados, esses padrões encontrados são representados na forma de regras ou subconjunto de características. Ainda segundo os autores a análise de agrupamento tem como objetivo encontrar grupos de observações relacionadas de forma que as observações que estejam no mesmo grupo sejam semelhantes entre si.

Por fim Tan, et al. (2009) relata que na classificação o objetivo é encontrar um modelo para predição de classes como função dos outros atributos. Para classificação o algoritmo *Random Forest* foi utilizado. Segundo Lorenzeti (2016) o algoritmo *Random Forest* constrói varias árvores de decisão usando um subconjunto aleatório de atributos obtidos do conjunto de dados originais, onde cada subconjunto gera uma árvore de decisão e para determinar a classe final de uma instancia verifica qual foi a classe mais escolhida dentre todas as árvores de decisão formada pelos subconjuntos, e este processo é repetido para todas as instancias. Ainda segundo os autores esse algoritmo apresenta resultados melhores do que com apenas uma árvore de decisão, porém, demanda mais tempo computacional.

Portanto, o objetivo deste trabalho foi aplicar o algoritmo *Random Forest* em imagens do sensor MODIS, utilizando o software estatístico R versão 3.3.1 (R Core Team, 2016), afim de mapear áreas de milho e soja no estado do Paraná para o ano-safra 2013/2014.

2. Material e Métodos

2.1. Área de estudo

A área de estudo abrangeu todo o estado do Paraná, (22°29' e 26°43'S; 48°2' e 54°38'W), (Figura 1). Para realizar o mapeamento utilizaram-se dados obtidos do sensor MODIS (Terra e Aqua) a partir dos produtos MOD13Q1 e MYD13Q1 que contêm o índice de vegetação da diferença normalizada (*Normalized Difference Vegetation Index* - NDVI), disponibilizados no banco de produtos da Base Estadual Brasileira (Esquerdo et al., 2010).

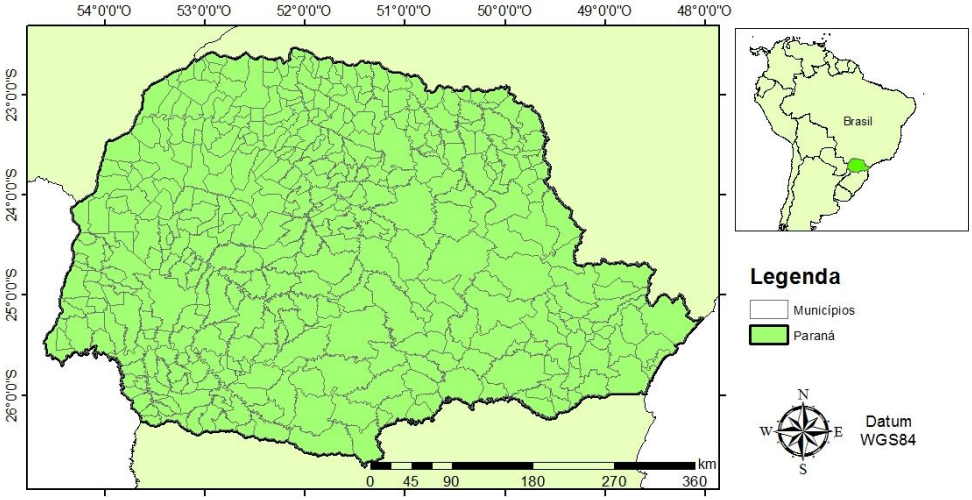

Figura 1. Localização do estado do Paraná.

2.2.Base de dados

As imagens coletadas do sensor MODIS possuem resolução espacial de 250 metros e resolução temporal de 8 dias (quando as plataformas TERRA e AQUA são combinadas), a série de imagens usadas neste trabalho compreendem o intervalo de 13 de agosto de 2013 a 15 de abril de 2014 totalizando 34 imagens. As datas de início e término foram determinadas de modo a abranger a época do plantio e colheita das culturas de verão (soja e milho). Na sequencia foi criada uma série temporal com as 34 imagens do sensor MODIS

2.3.Conjunto de teste

Na sequencia foi construído um conjunto de teste, onde primeiramente foram coletados pixels puros em imagens do sensor *Operational Land Imager* (OLI) do satélite Landsat8 (INPE, 2016) na composição de falsa cor RGB-564 com uso do software Arcgis, utilizando um grid de 250m de MODIS com o intuito de extrair pixels puros que possa ser convertido para uma cena MODIS.

Para cobrir toda a área de estudo foram necessárias 14 cenas, sendo que em cada uma delas foram adquiridas amostras de milho, soja, água, cidade e floresta, de modo a separar os alvos em três classes, sendo elas: milho, soja e outros.

2.3 Aplicação do algoritmo Randon Forest

Seguindo a ideia do KDD os processos realizados anteriormente estão vinculados as fases de coleta de dados, pré-processamento e formatação, após realizado essas etapas se da inicia a fase mais importante do KDD, a mineração de dados.

Realizou-se essa fase no software R versão 3.3.1 (R Development Core Team, 2016) com os pacotes *raster* (Hijmans, 2016); *caret* (Kuhn et al, 2016) e *rgdal* (Bivant et al., 2016) e foi utilizado como variáveis de entrada a serie temporal e em seguida foi vinculado as informações de cada pixel dessa serie com o conjunto de teste. Em seguida foi aplicado o algoritmo *Random Forest*. O próprio algoritmo apresenta valores de exatidão global e kappa.

Com isso gerou-se uma classificação para todo o Paraná separando em três classes milho, soja e outros. Por fim a ultima etapa do KDD é a verificação, nesta fase foi verificada acurácia do mapeamento gerado através do índice kappa, índice de exatidão global (Congalton, 1991; Congalton e Green, 1999) e na sequencia foi feita a comparação da área total plantada de cada cultura com dados da Companhia Nacional de Abastecimento (CONAB).

3. Resultados e Discussões

Após a aplicação do KDD, gerou-se a classificação de soja e milho para o estado do Paraná no ano safra de 2013/2014 obtida pelo algoritmo Random Forest (Figura 2).

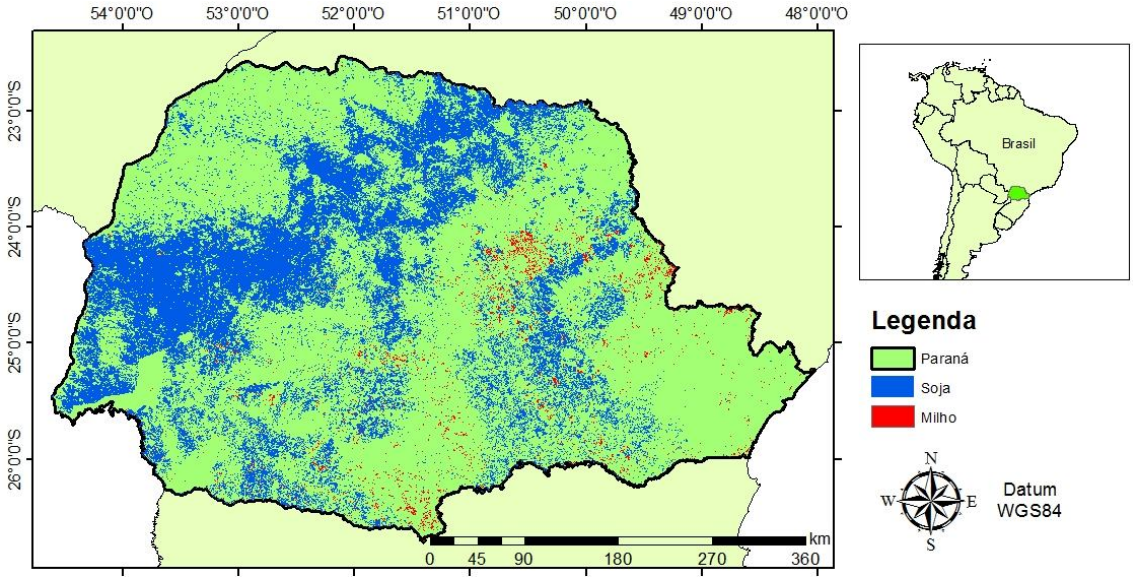

Figura 2. Classificação de culturas de verão no estado do Paraná ano-safra 2013/2014

O mapeamento (Figura 2) evidencia a existência do "cinturão da soja", que é uma faixa que se desloca da região Oeste até a região Norte do estado do Paraná onde a cultura da soja é predominante. De acordo com essa classificação extraiu-se as áreas totais de cada cultura que apresenta os valores da área cultivada de milho e soja em hectares (ha) obtida com o algoritmo (Tabela 1).

Tabela 1 Comparação de áreas cultivadas de culturas de verão com dados da CONAB

Cultura	*Valor obtido pelo método (ha)*	*Valor disponibilizado pela CONAB (ha)*
Milho	336.100	668.200
Soja	6.199.400	5.019.000

Observa-se que a classificação do milho gerou um resultado 49,7% menor que o valor apresentado pela CONAB (Tabela 1), enquanto que para a soja o valor da classificação foi 23,51% maior que o valor comparado com dados da CONAB.

O ideal seria que essas diferenças fossem as menores possíveis, entretanto as maiores vantagens dessa metodologia é o baixo custo (uso de imagens gratuitas), rapidez e objetividade que segundo Duft et al. (2011) não é possível quando é utilizada as metodologias dos órgãos oficiais.

Referente ao índice kappa essa classificação obteve 0,9678 o que é considerado excelente de acordo com Landis e Koch (1977). Para exatidão global o valor foi de 98,61% o que também representa um ótimo resultado, mostrando com isso o potencial dessa metodologia para o mapeamento de culturas de verão.

Os valores obtidos possuem relevância maior se comparado com trabalhos que utilizam o mesmo sensor para tarefas semelhante, como o caso de Lamparelli et al. (2008) que utilizou imagens MODIS/ e Landsat 5 para mapear soja no estado do Paraná obteve o índice Kappa entre 0,6 e 0,8. Já no trabalho de Johann (2012), cujo o objetivo foi mapear áreas da cultura de verão no estado do Paraná, obteve índice kappa e exatidão global de 0,8945 e 94,72%

respectivamente. Enquanto que no trabalho de Zhong et al. (2016) os índices kappa e exatidão global foram de 0,804 e 87,2% respectivamente.

4. Conclusões

A metodologia proposta nesse trabalho obteve excito no mapeamento de soja e milho no estado do Paraná para o ano safra 2013/2014, pois o método é rápido e de custo baixo, tendo em vista que tanto as imagens quanto o software R são gratuitos.

Além disso, o método se destaca por possuir excelentes valores de kappa e índice de exatidão global, porém, os resultados obtidos do mapeamento subestimaram em aproximadamente 50% a área de milho se comparado com dados oficiais e superestimou a área de soja em cerca de 24%, indicando que o método deve ser melhorado nesse aspecto.

Contudo a metodologia proposta mostra-se promissora, entretanto deve ser trabalhada a preparação das imagens com o intuito de diminuir a diferença entre a quantidade de área obtida no método e os dados oficiais.

5. Agradecimentos

Ao Programa de Pós-graduação Stricto Sensu em Engenharia Agrícola (PGEAGRI) pela oportunidade bem como ao Laboratório de Estatística Aplicada (LEA) da UNIOESTE/Campus Cascavel, pela infraestrutura disponibilizada para realizar este trabalho.

6. Referências

Brown, J. C.; Kastens, J. H.; Coutinho, A. C.; Victoria, D. Bishop, C. R. Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. **Remote Sensing of Enviroment**. v. 130, p.39-50, 2013.

CONAB, Companhia Nacional de Abastecimento. **Acompanhamento safra brasileira de grãos**, v. 5- Safra 2015/16 - Quinto levantamento, Brasília, p. 1-182, 2016.

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

Congalton, R.G.; Green, K. **Assessing the accuracy of remotely sensed data: principles and practices**. Boca Raton: CRC Press, 1999. 160p.

Duft, D. G; Johann, J. A; Rocha, J. V; Lamparelli, R A. C. "Metodologia para geração de mascaras de cultura de verão para o ano safra 200/2008 no estado do Paraná por meio de índice de vegetação do sensor MODIS" in: **Anais XV Simpósio Brasileiro de Sensoriamento Remoto** – SBSR, Curitiba-PR, 30 de abr. a 05 de maio de 2011 pag. 140 - 147

Esquerdo, J. C. D. M.; Antunes, J. F. G.; Andrade, J. C. de. **Desenvolvimento do banco de produtos MODIS na Base Estadual Brasileira**. (Comunicado Técnico, 100) - Campinas: Embrapa Informática Agropecuária, 7 p., 2010.

ESRI - ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. **Products**. Disponível em: <http://www.esri.com/>. Acesso em: 4 nov. 2016.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthrusamy, R. **Advances in Knowledge Discovery & Data Mining**. California: AAAI/MIT, 1996.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Banco de Dados Agregados**: Sistema IBGE de Recuperação Automática – SIDRA. 2016. Disponível em: <http://www.sidra. ibge.gov.br>. Acesso em: 24 fev. 2016.

INPE – Instituto Nacional de Pesquisas Espaciais. **Catálogo de imagens**. 2016. < http://www.dgi.inpe.br/CDSR/>. 18 Abr. 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.

Johann, J. A.; Rocha, J. V.; Duft, D. G.; Lamparelli, R. A. C. Estimativa de áreas com culturas de verão no Paraná, por meio de imagens multitemporais EVI/Modis. **Pesquisa Agropecuaria Brasileira**, n. 9, v. 47, p. 1295-1306, 2012.

Lamparelli, R.A.C.; Carvalho, W.M.O. de; Mercante, E. Mapeamento de semeaduras de soja (Glycine max (L.) Merr.) mediante dados MODIS/Terra E TM/Landsat 5: um comparativo. **Revista Engenharia Agrícola**, v.28, p.334 - 344, 2008.

Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, p.159-174, 1977.

Lorenzett, C. D. C; Telöcken, A. V. Estudo Comparativo entre os algoritmo de mineração de dados Random Forest e J48 na tomada de decisão. In: **Anais do II Simpósio de Pesquisa e Desenvolvimento em Computação.** Cruz alta – RS, de 09 a 13 de maio de 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. https://CRAN.R-project.org/package=caret

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

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robert J. Hijmans (2016). raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster

Santos, J. S.; Fontana, D. C.; Silva, T. S. F.; Rudorff, B. F. T. Identificação da dinâmica espaço-temporal para estimar área cultivada de soja a partir de imagens MODIS no Rio Grande do Sul. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, PB, v.18, n.1, p.54-63, 2014.

Tan, P. N; Steinbach, M; Kumar, V. **Introdução ao DATA MINING Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

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.