Capítulo 4

Big Earth Observation Data e aprendizado de máquina para mapeamento da agricultura sustentável no Brasil

Patrick Calvano Kuchler
Margareth Simões
Agnés Bégué
Rodrigo Demonte
Damien Arvor

Resumo: a implementação do iLP, ou seja, a diversificação, rotação, consorcio e/ou sucessão das atividades agrícolas e de pecuária na mesma área formando um único sistema, é considerada uma importante estratégia de intensificação agrícola sustentável para brasil, com diversos impactos positivos com destaque na conservação do solo e rentabilidade e viabilidade econômica. O acompanhamento da implantação desta iniciativa é fundamental como instrumento de gestão pública, porém ainda é um desafio. Nesta direção, este trabalho discute a aplicação dos conceitos de BIG DATA e aprendizado de máquina para o sensoriamento remoto. Como teste foi utilizado o classificador *Randon Forest* (RF) aplicado a séries temporais MODIS para analisar a capacidade de detecção de certos iLPs. Para isso, avaliamos a precisão do RF aplicado ao NDVI do MODIS para os anos de 2012 a 2016 em uma área no norte do mato grosso. Dois modelos foram testados: (i) usando 11 métricas fenológicas derivadas do MODIS (ii) usando as métricas e os dados originais. O índice kappa para (i) foi de 0,63, sendo 9 deles com potencial discriminatório; o resultado de (ii) foi de 0,84, onde somente 01 métrica foi significativa para discriminação. Nossos resultados indicam que o uso da técnica de classificação RF com dados MODIS tem grande potencial para compor uma metodologia de monitoramento do iLP, sendo o grande desafio o tratamento das SITS em larga escala, necessitando em termos de arquitetura de sistemas processamento paralelo em nuvem, tecnologia esta, cada vez mais disponível.

Palavras-chave: sensoriamento remoto, séries-temporais, *random forest*, Mato Grosso, sistemas integrados, Big Earth Observation Data, aprendizado de máquina.

1. INTRODUÇÃO

Atualmente o Brasil é uma das maiores potências agrícolas no mercado mundial, exercendo papel significativo no suprimento atual e das próximas décadas na demanda global de alimentos e energia (Arvor et al., 2012). A dinâmica atual de expansão de áreas agropecuárias fez com que o Brasil entrasse como 6° maior país emissor de Gases de Efeito Estufa (GEE) reportado na 17ª Conferência das Partes (COP-17). Em um esforço para reverter e frear esta situação, o antigo governo brasileiro incentivou a adoção de medidas para promover práticas agrícolas que possam intensificar a produção de forma menos agressiva ao ambiente. Estas medidas são fruto de um compromisso assumido voluntariamente de redução da emissão de GEE entre 36,1% e 38,9% durante a COP15 ocorrida em Copenhague no ano de 2009. Neste momento, foram propostas algumas ações, dentre as quais, na área agrícola, a promoção da agricultura de baixo carbono, dando origem ao plano setorial de mitigação das mudanças climáticas para a agricultura, o chamado plano ABC. Os Sistemas Integrados (SI) merecem ser destacados neste contexto como uma estratégia muito promissora para atingir as metas firmadas. Combinando culturas, pecuária e / ou silvicultura nas mesmas áreas eles formam um único sistema harmônico capaz de aumentar a fertilidade e a matéria orgânica contida no solo, que favorece a produção de biomassa (Bungenstab, 2012; Carvalho et al., 2014). Os SI podem ser divididos em dois grandes grupos: A integração Lavoura, Pecuária e Floresta (iLPF) e a integração Lavoura e Pecuária (iLP). Neste trablho iremos focar no iLP, que é mais amplamente implementado e é baseado em consorcio, sucessão de culturas e/ou rotação, onde sempre terá que haver o elemento pastagem. O iLP ainda pode ser definido como Intra-Anual, quando a integração é realizada no mesmo ano/safra e a Inter-Anual, quando a integração é realizada em anos diferentes. Uma estimativa de implantação de iLP Inter-Anual utilizando Análise Espacial de dados oriundos de SR é encontrado em Kuchler et al., 2019.

Um ponto em destaque no plano ABC se refere ao desafio de criar mecanismos efetivos para monitorar e acompanhar o desenvolvimento dessas ações propostas (MAPA, 2011).

As séries temporais de imagens de satélite (SITS) são amplamente utilizadas para o mapeamento de sistemas agrícolas e têm grande potencial para compor uma metodologia de monitoramento do referido plano. Neste sentido, para monitorar o crescimento das culturas ao longo das estações e identificar mudanças nas práticas agrícolas, faz-se necessária uma coleção de dados capaz de construir uma série temporal de Índice de Vegetação (IV), por este motivo satélites com alta resolução temporal são instrumentos necessários principalmente em regiões com alta incidência de nuvens. Dados do satélite MODIS, oferecem a vantagem de uma revisita diária, possibilitando a realização de interpolações temporais e a geração de produtos com poucas nuvens. O processamento de uma série de imagens em grandes porções de terra aplicado a uma série histórica não é trivial considerando métodos clássicos de classificação. Os conceitos de *Big Data* em conjunto com os algoritmos de aprendizagem de máquina vem suprindo cada vez mais estes desafios dentro da área de SR e *Earth Observation*, porém impondo outros como é o caso da disponibilidade necessária de amostras de campo para treinamento dos algoritmos.

2. MATERIAIS E MÉTODOS

Mato Grosso (MT) é um estado brasileiro dentro da “Arco do desmatamento”, onde a agricultura está se expandindo rapidamente contribuindo para o aumento da pressão global da terra e mudança do uso da terra. Considerando que Mato Grosso é o principal produtor de gado e soja do país e fica adjacente à porção mais densa da floresta amazônica, a adoção de SI neste estado poderia ajudar a alcançar o desenvolvimento de tecnologias mais eficientes e sustentáveis. Foi selecionada uma área na porção norte do estado para testar a metodologia proposta, onde havia disponibilidade de dados. figura 1 apresenta a localização da área de estudo.

Correntes atuais assumem que para o monitoramento da agricultura por SR, uma única imagem de satélite, de uma única data, não fornece informações espectrais suficientes para identificar culturas plantadas em uma determinada estação, por este motivo, séries temporais de dados de Índices de Vegetação (IV) são utilizados para o mapeamento de sistemas agrícolas. Este tipo de abordagem utilizando dados do sensor MODIS estão sendo testadas no Brasil e em particular no MT para o mapeamento de culturas anuais.

Figura 01: Localização Área de Estudo

Spera et al. (2014) usam o SR para examinar padrões de terra cultivável para a expansão agrícola no estado na janela temporal de 2001 a 2011. Eles usaram uma série temporal de EVI do MODIS, com um algoritmo de árvore de decisão. Foram identificados dados de cultivos específicos para mapear cinco classes: soja, algodão, soja-milho, soja-algodão, e culturas irrigadas.

Arvor et al. (2011) também utilizou a série temporal MODIS EVI para identificar cinco classes de culturas: culturas de soja, milho e algodão plantadas em sistemas de cultivo simples ou duplo. Eles assumem que o milho só é plantado em consórcio com soja. Os autores aplicam uma segmentação para produzir resultados mais homogêneos. A precisão relatada é de 85% para a máscara de agricultura e 74% para a classificação.

Para descrever a dinâmica espacial da produção agrícola no MT de 2001 a 2014, Kastens et al. (2017) usam a série temporal MODIS NDVI. Eles utilizam dados de referência de campo de 2009 a 2016 para criar uma base de aprendizagem do classificador RF. A precisão relatada foi de 79% para distinguir classes de culturas (soja-pousio, algodão em pousio, soja-algodão e soja). Chen et al. (2018) desenvolveram uma metodologia para identificar tipos de culturas, incluindo soja, algodão e milho em sistemas: soja-milho, soja-algodão, soja-pasto, soja-pousio, pousio-algodão e cultura única no MT. Foram utilizados dados NDVI do MODIS que passaram por processo de filtragem e extração de métricas fenológicas das séries temporais pré-processadas e posteriormente foi utilizado um classificador de árvore de decisão para o mapeamento nos anos de 2015 e 2016. Eles alcançaram uma acurácia de 90% para áreas cultivadas, 73% para padrões de culturas e 86% para tipos de culturas.

Foram publicados no mês de novembro de 2017 pelo portal Pangaea dados do mapeamento de tipos de cultura entre os anos de 2001 e 2016 no estado do MT com as seguintes classes e seus respectivos coeficientes de acurácia: Cerrado (99%), Pousio-Algodão (100%), Floresta (99%), Pasto (95%), Soja-Milho (87%), Soja-Algodão (99%), Soja-Pousio (100%), Soja-Milheto (84%), e soja Girasol (85%) (CÂMARA et al., 2017).

Entre os estudos citados, é possível identificar a ausência de métodos para a detecção de iLP. Para o nosso trabalho, foi utilizada uma série temporal com intervalo de 16 dias do índice de vegetação de diferença normalizada (NDVI, sigla em inglês) do MODIS MOD13Q1 para os anos de 2012 a 2016 totalizando 155 imagens.

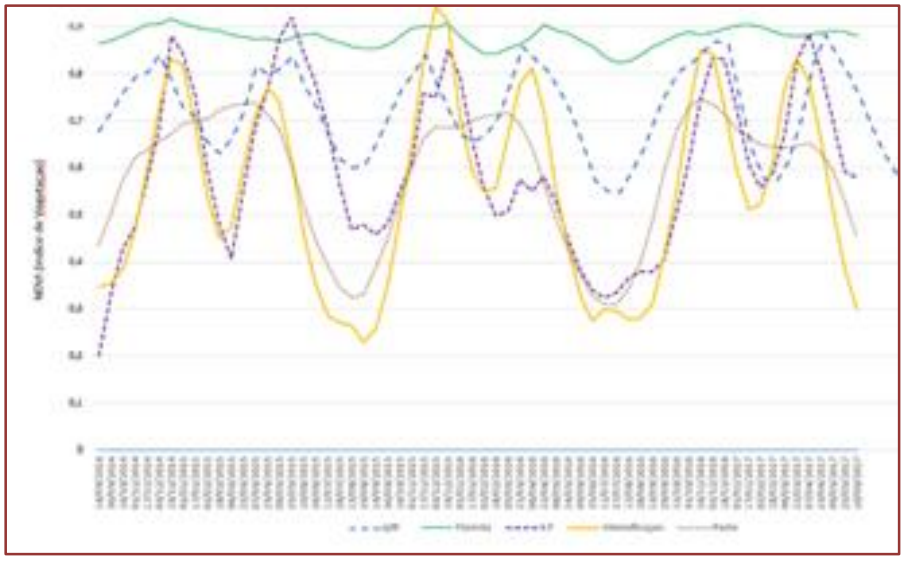
Jönsson e Eklundh (2004) desenvolveram o programa TIMESAT que tem com o objetivo de extrair informações sobre a fenologia da vegetação com base na sazonalidade através de dados de IV oriundos de produtos obtidos por séries temporais de SR, que inicialmente devem passar pelo pré-processamento de filtragem e suavização das curvas. o TIMESAT extrai 11 diferentes atributos da série temporal, criando arquivos matriciais do que é chamado de métricas fenológicas. São elas: a- início e b-fim do ciclo fenológico, c e d- o nível de 80% de distribuição dos dados a direita e a esquerda, e- ponto mais alto, f- amplitude sazonal, g- a duração da estação, h e i- integrais do valor cumulativo de crescimento da vegetação j- o valor mais baixo e k- é o meio do ciclo fenológico, Jönsson e Eklundh (2012).

Para a filtragem e suavização da série temporal NDVI MODIS, foi utilizado o método *Savitsky Golay* (SG) através do software TIMESAT 3.2, indicado como melhor método por Chen et al., 2004. Posteriormente, 11 métricas fenológicas foram extraídas com valores relativos a informação de sazonalidade da vegetação, resultando em 55 dados matriciais.

Realizamos experimentos na mesma região do estado para avaliar o nível adequado de pré-processamento para um ano de séries temporais MODIS no intuito de mapear os SI, o estudo e seus resultados podem ser encontrados em Kuchler et al., 2020.

Além de dados de SR, foram utilizados dados da fazenda Gamada, URT (Unidade de Referência Tecnológica da EMBRAPA) e mais dados de terreno coletados via aplicações SatVeg, TerraClass e dados *Pangae*. Com estes dados e informações do terreno, foi construída uma base de aprendizagem do algoritmo RF. A junção dos Ao todo, foram selecionadas 197 amostras de 05 classes diferentes, sendo iLP (36 amostras), iLPF (1 amostra), Floresta (57 amostras), Pasto (60 amostras), Agricultura Intensificada, que representa duas culturas de verão (43 amostras). No local selecionado para o estudo, não foram encontradas mais amostras de iLPF além da fazenda Gamada, por este motivo a classe iLPF foi eliminada do modelo final. A figura 3 apresenta exemplo de curva padrão para cada uma das 5 classes extraída dos dados suavizados pelo método SG, representando o valor médio das parcelas da fazenda Gamada.

Figura 03: Perfis das amostras por classe



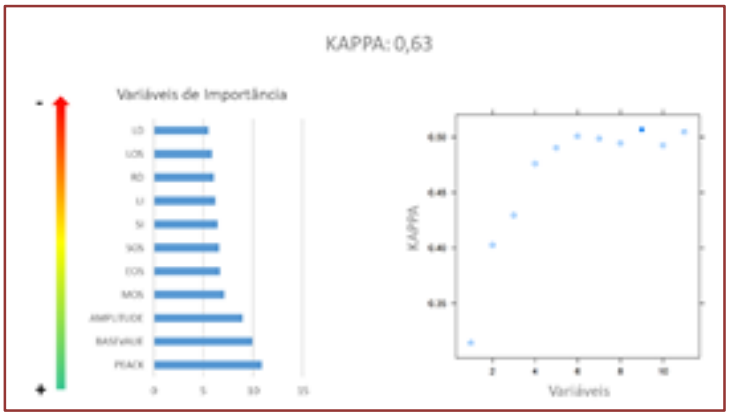
Randon Forest (RF) é uma técnica de aprendizado de máquina que gera uma infinidade de árvores de decisão aleatórias que são agregadas, para então gerar uma classificação (BREIMAN, 2001). Cada árvore de classificação (500 árvores em uma floresta típica) é construída de um conjunto amostrado aleatoriamente composto por aproximadamente um terço do conjunto completo de dados, para uma boa aprendizagem do modelo, é necessária uma quantidade significativa de amostras (CUTLER, 2007). Em estudos de classificação de uso e ocupação do solo, o classificador é considerado estável e relativamente eficiente, além de envolver poucos parâmetros definidos pelo usuário e mesmo assim gerar bons níveis gerais de precisão (LAWRENCE, 2006). Dos resultados gerados através da abordagem de RF, a diminuição média na precisão (DMP) para uma variável permite avaliar a importância de cada variável usada para classificação. Quanto mais a precisão do RF diminui devido à exclusão de uma única variável, mais importante é essa variável. Consequentemente, valores mais altos de DPM indicam variáveis que são mais importantes para a classificação (CUTLER, 2007).
Como dados de entrada para o classificador RF, foram utilizados os dados originais do MODIS, assim como os dados das métricas fenológicas.

3. RESULTADOS

Dois modelos de classificação RF foram aplicados, (i) utilizando somente as métricas fenológicas e (ii) utilizando as métricas e a série temporal original. O índice *kappa* para (i) foi de 0,63 sendo que das 11 métricas, 09 apresentam potencial discriminatório entre classes. Neste resultado, as três variáveis que obtiveram maior valor DPM foram o ‘valor de pico’ (PEACK), ‘valor de base’ (BASEVALUE) e ‘amplitude’ (AMPLITUDE),

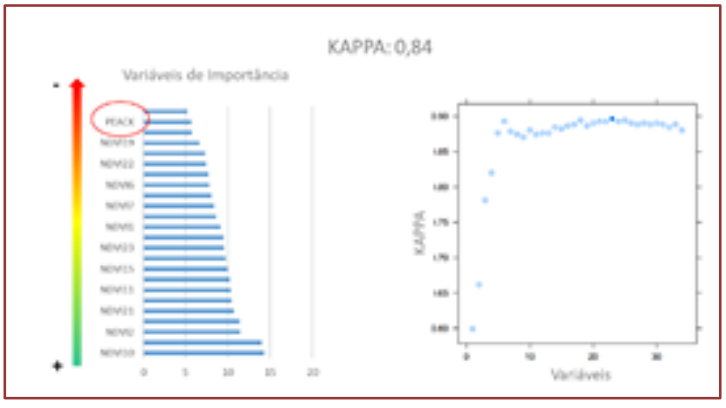
respectivamente. Os menores valores DPM foram a ‘derivativa a esquerda’(LD), a ‘duração da estação’ (LOS) e a ‘derivativa a direita’ (RD), respectivamente. O algoritmo indica as variáveis que não tem representatividade para a discriminação entre as classes, que neste caso foram as duas últimas, ou seja, a ‘derivativa a direita’ (RD) e a ‘duração da estação’(LOS). A figura 4 apresenta os gráficos gerados pelo algoritmo.

Figura 04: A) Valores DPM das variáveis. B) Número de variáveis efetivamente discriminativas.Perfis das amostras de cada classe na fazenda Gamada



O segundo modelo utilizando os dados originais e das métricas apresentou um valor de *kappa* de 0,89. Os valores de DPM apresentaram uma maior significância nas variáveis do NDVI bruto, excluindo praticamente todas as métricas fenológicas, considerando somente o ‘valor de pico’ (peack). A figura 5 apresenta os gráficos gerados pelo algoritmo para esta classificação.

Figura 6: A) Valores DPM das variáveis. B) Número de variáveis efetivamente discriminativas



4. DISCUSSÃO

Devido ao baixo valor de acurácia encontrado, pretendeu-se realizar uma classificação utilizando os dois pacotes de dados para analisar se há alguma melhora na classificação e também para comparar os valores de DPM entre os dados brutos e das métricas geradas. O resultado melhorou consideravelmente, atingindo um valor de *kappa* bem acima. Os valores de DPM apresentaram uma maior significância nas variáveis do NDVI bruto, excluindo praticamente todas as métricas fenológicas.

5. CONCLUSÕES

Os estudos atuais sobre a identificação e o mapeamento da dinâmica agrícola no estado do MT apresentam diversas abordagens no que se refere a técnicas e objetivos, porém a respeito de dados de sensores remotos óticos, ainda há uma grande tendência de esgotar a aplicação dos produtos MODIS, pois a região em que o estado se localiza é de alta frequência de nuvens, dificultando outros sensores óticos a atingir a

temporalidade desejada. A limitação da resolução espacial não se caracteriza como um impeditivo para a aplicação dos estudos de agricultura nesta região, visto que as glebas são geralmente acima de 10 ha.

Nossos resultados apontam que a utilização de métricas fenológicas não representaram melhora significativa na detecção de SI no MT. O resultado baseado nas métricas tem um valor considerado baixo quando comparado com o resultado do modelo que utiliza as séries originais. Quando integrada a série original com as métricas, a melhora no valor de *kappa* é alcançada, porém sem grande representatividade. Os resultados apontam que a utilização da técnica de classificação RF em abordagem multitemporal tem grande potencial para compor uma metodologia de monitoramento dos SI. No caso apresentado, foram utilizadas 155 imagens para uma pequena área no MT. A aplicação de um modelo que cubra todo o estado ou todo o corredor de grãos do país, alcançando estados como Goiás, Mato Grosso, Mato Grosso do Sul, Tocantins, Pará resularia em um processamento na ordem de terabytes. As técnicas de aprendizado de máquina e *Big Earth Observation Data* são essenciais em momentos como este, onde a grande disponibilidade de dados não são mais possíveis de serem processadas de forma tradicional, ao mesmo tempo, sistemas de computação de alto desempenho com processamento em nuvem estão se tornando cada vez mais disponíveis, muitas vezes sendo comercializadas. Estes conceitos lançam uma nova dimensão e escala sobre o sensoriamneto remoto, mas também o desafiam a medida que amostras de "verdade de campo" são excenssiais para a construção uma base de aprendizagem significativa para alcançar uma boa acurácia na classificação. A limitação atual não é mais referente a área a ser mapeada, que pode ser continental ou mesmo planetária, mas é referente a quantidade de amostras necessárias para comopor a base de aprendizagem dos algorítmos de aprendizado de máquina.

AGRADECIMENTOS

Esse trabalho foi realizado no âmbito do projeto CAPES-COFECUB GeoABC e do projeto Europeu H2020-MSCA-RISE-2015 ODYSSEA project (Project Reference: 691053), EMBRAPA Agrosilvipastoril e Embrapa Labex Europa.

REFERÊNCIAS

[1] Arvor D, Jonathan, M.; Meirelles, M. S. O. P.; Dubreuil, V.; Durieux, L. "Classification of MODIS EVI timeseries for crop mapping in the state of Mato Grosso, Brazil". International Journal of Remote Sensing, v. 32, n 22, pp. 7847 – 7871, 2011.

[2] Bungenstab, D.J. "Sistemas de Integração Lavoura-Pecuária-Floresta – A Produção Sustentável". ed. EMBRAPA, Brasília, 2012.

[3] Brasil. Ministério da Agricultura. "Plano Setorial de mitigação e adaptação ao clima, Livestock and Food Supply". Brasilia-DF: MAPA, 2011.

[4] Ministério da Ciência, Tecnologia e Inovação. "Estimativas anuais de emissões de gases de efeito estufa no Brasil". Brasília-DF: MCTI, 2014.

[5] Câmara, Gilberto; Picoli, Michelle; Simoes, Rolf; Maciel, Adeline; Carvalho, Alexandre; Coutinho, Alexandre; Esquerdo, Julio; Antunes, Joao; Begotti, Rodrigo; Arvor, Damien. "Land cover change maps for Mato Grosso State in Brazil: 2001-2016". links to files. PANGAEA, https://doi.org/10.1594/PANGAEA.881291, 2017.

[6] Carvalho, J.L.N., Raucci, G.S., Frazao, L.A., Cerri, E.C., Bernoux, M., Cerri, C.C. "Crop-pasture rotation: a strategy to reduce soil greenhouse gases emissions in the Brazilian Cerrado". Agric. Ecosyst. Environ. v. 183 n 1, pp. 167–175. 2014

[7] Chen, Y., Dengsheng, L., Emilio Moran, Mateus Batistella, Luciano Vieira Dutra, Ieda Del'Arco Sanches, Ramon Felipe Bicudo da Silva, Jingfeng Huang, Alfredo José Barreto Luiz, Maria Antonia Falcão de Oliveira. "Mapping croplands, cropping patterns, and crop types using MODIS timeseries data". Int J Appl Earth Obs Geoinformation. v. 66 pp. 133-147. 2018.

[8] Chen, J.; Jönsson, P.; Tamura, M.; Gu, Z.; Matsushita, B.; Eklundh, L. "A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter". Remote Sensing of Environment, v. 91, n 3, pp. 332 – 344. 2004.

[9] Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. 'Random forests for classification in ecology'. Ecology, v.88, pp. 2783–2792. 2007

[10] Jönsson, P.; Eklundh, L. "TIMESAT – a program for analyzing time-series of satellite sensor data". Computers and Geosciences, v. 30, n 8, pp. 833 – 845. 2004

[11] Kastens, J., Brown, J., Coutinho, A., Bishop, C., Esquerdo, J. “Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil”. PLOS ONE 12 (4), e0176168. 2017.

[12] Kuchler, P.C.; Bégué, A.; Simões, M.; Gaetano, R.; Arvor, D.; Ferraz, P.D.R “Assessing the optimal preprocessing steps of MODIS time series to map cropping systems in Mato Grosso, Brazil”. International Journal of Applied Earth Observation and Geoinformation 92 (outubro de 2020): 102150. https://doi.org/10.1016/j.jag.2020.102150.

[13] Kuchler, P.C.; Bégué, A.; Simões, M.; Arvor, D.; Ferraz, P.D.R “SENSORIAMENTO REMOTO E ANÁLISE ESPACIAL: UMA CONTRIBUIÇÃO PARA O MAPEAMENTO DOS SISTEMAS INTEGRADOS DE PRODUÇÃO AGROPECUÁRIA”. In Aplicações e Princípios do Sensoriamento Remoto 3, 1–10. Atena Editora, 2019. https://doi.org/10.22533/at.ed.3791923091.

[14] Lawrence, R.L.; Wood, S.D.; Sheley, R.L. “Mapping invasive plants using hyperspectral imagery and breiman cutler classifications (randomforest)”. Remote Sens. Environ. v.100, pp. 356–362. 2006.

[15] Spera, S. A., Cohn, A. S., VanWey, L. K., Mustard, J. F., Rudorff, B. F.,Risso, J., Adami, M. “Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics”. Environmental Research Letters v.9 n.6, 064010. 2014.