

# Desempenho do Algoritmo de Classificação de Imagens Random Forest para Mapeamento do Uso e Cobertura do Solo no Cerrado Brasileiro

*Performance of the Random Forest Image Classifier for Mapping Land Use and Land Cover in the Brazilian Cerrado*

David Fernando Cho<sup>1,3</sup> , Samuel Fernando Schwaida<sup>2,3</sup> , Rejane Ennes Cicerelli<sup>3</sup> , Tati Almeida<sup>3</sup> , Ana Paula Marques Ramos<sup>4</sup>  & Edson Eyji Sano<sup>3,5</sup> 

<sup>1</sup>Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais, SCEN Ibama, Brasília, DF, Brasil

<sup>2</sup>Ministério do Meio Ambiente, Esplanada dos Ministérios, Brasília, DF, Brasil

<sup>3</sup>Universidade de Brasília, Instituto de Geociências, Campus Universitário Darcy Ribeiro, Brasília, DF, Brasil

<sup>4</sup>Universidade do Oeste Paulista – UNOESTE, Pós-Graduação em Meio Ambiente e Desenvolvimento Regional, Presidente Prudente, SP, Brasil

<sup>5</sup>Embrapa Cerrados, Planaltina, DF, Brasil

E-mails: [samuelschwaida@gmail.com](mailto:samuelschwaida@gmail.com); [davidfcho@gmail.com](mailto:davidfcho@gmail.com); [rejaneig@unb.br](mailto:rejaneig@unb.br); [tati\\_almeida@unb.br](mailto:tati_almeida@unb.br); [anamos@unoeste.br](mailto:anamos@unoeste.br); [edson.sano@ibama.gov.br](mailto:edson.sano@ibama.gov.br)

## Abstract

The Cerrado is a highly diversified ecosystem and provides habitat for many species, however, it has undergoing marked degradation in recent decades due to the expansion of agricultural commodity production. This scenario reinforces the need for continuous monitoring of land use and land cover (LULC) changes, whether with a focus on environmentally sustainable agricultural production or market understanding. Recently, machine learning algorithms have become a promising and innovative approach to remote sensing data processing. Thus, this study aimed to evaluate the potential of the Random Forest image classification algorithm for LULC mapping and classification in the Brazilian Cerrado. The selected study area was the municipalities of Natividade, Chapada da Natividade, and São Valério da Natividade, located in the state of Tocantins. The basic materials of this study were the digital elevation model produced by the Shuttle Radar Topography Mission (SRTM), the night light images obtained by the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor onboard the Suomi National Polar-Orbiting Partnership (Suomi NPP) and NOAA-20 satellites, and the Landsat 8 Operational Land Imager satellite (OLI) multispectral images acquired from May to October 2013. All analyzes were performed on the Google Earth Engine platform that allows cloud computing. A cube of images was generated containing 38 layers that were classified by the Random Forest algorithm, with 500 decision trees. For the algorithm training, random points from each LULC class mapped by the TerraClass Cerrado 2013 project were used. Considering the TerraClass Cerrado 2013 mapping as the ground truth, a Kappa index of 0.64 was obtained. There was a significant overestimation of annual cropland and urban areas. The proposed methodology presented a good potential for less expensive and less time demanding LULC mapping of the Cerrado.

**Keywords:** Geotechnologies; Machine learning; Cloud processing

## Resumo

O Cerrado é um ecossistema altamente diversificado e fornece habitat para muitas espécies, porém, vem sofrendo degradação acentuada nas últimas décadas devido à expansão da produção de *commodities* agrícolas. Esse cenário reforça a necessidade de contínuo monitoramento das mudanças de uso e cobertura do solo, seja com foco na produção agrícola ambientalmente sustentável ou no entendimento do mercado. Recentemente, os algoritmos de aprendizagem de máquina têm-se concretizado como uma abordagem promissora e inovadora para processamento de dados de sensoriamento remoto. Assim, esse trabalho teve por objetivo avaliar o potencial do algoritmo de classificação de imagens *Random Forest* para o mapeamento e classificação do uso e cobertura do solo no Cerrado Brasileiro. A área de estudo selecionada foram os municípios de Natividade, Chapada da Natividade e São Valério da Natividade, localizados no estado do Tocantins. Os materiais básicos deste estudo foram o modelo digital de elevação produzido pela missão *Shuttle Radar Topography Mission* (SRTM), as imagens de luzes noturnas obtidas pelo sensor *Visible Infrared Imaging Radiometer Suite* (VIIRS) dos satélites *Suomi National Polar-Orbiting Partnership* (Suomi NPP) e NOAA-20 e as imagens multiespectrais do satélite Landsat 8 Operational Land Imager (OLI), adquiridas entre os meses de maio a outubro de 2013. Todas as análises foram realizadas na plataforma Google Earth Engine que permite processamento de dados em nuvem. Foi gerado um cubo de imagens contendo 38

camadas que foram classificadas pelo algoritmo Random Forest, com 500 árvores de decisão. Para o treinamento do classificador, foram utilizados pontos aleatórios de cada classe de mapeamento do projeto TerraClass Cerrado 2013. Considerando o mapeamento do TerraClass Cerrado 2013 como verdade terrestre, obteve-se um índice Kappa de 0,64. Houve superestimação expressiva da agricultura anual e da área urbana. A metodologia proposta apresentou um bom potencial como uma alternativa de menor custo e tempo para o mapeamento do uso e cobertura do solo do Cerrado..

**Palavras-chave:** Geotecnologias; Machine learning; Processamento em nuvem

## 1 Introdução

O Cerrado é o segundo maior bioma brasileiro, é considerado como *hotspot* para conservação da biodiversidade e é um importante fornecedor de serviços ecossistêmicos que trazem benefícios para as populações humanas que dependem diretamente e indiretamente desse bioma como fonte de alimento, água, materiais e polinizadores (Alencar et al. 2020). Apesar disso, aproximadamente metade da sua área original foi convertida para atividades produtivas nos últimos 45 anos e menos de 9% do bioma encontra-se sob algum tipo de proteção integral (Brasil 2015; Klink & Machado 2005; Vieira et al. 2018). Em função da forte pressão antrópica sobre o bioma, o mapeamento e monitoramento das mudanças do uso do solo são fundamentais para traçar estratégias e políticas de proteção do bioma e sua biodiversidade, bem como para o ordenamento do território e desenvolvimento econômico.

Diante desse cenário de rápida e intensa mudança do uso e cobertura do solo no Cerrado, é fundamental conhecer a dinâmica desta transformação de modo rápido e preciso por meio do uso de ferramentas eficientes (Alencar et al. 2020). Os principais esforços de mapeamento do Cerrado realizados em escala de bioma são o Projeto de Conservação e Utilização Sustentável da Diversidade Biológica Brasileira (Probio) (Sano et al. 2007, 2019) e o Mapeamento do Uso e Cobertura Vegetal do Cerrado - TerraClass Cerrado (Brasil 2015), ambas iniciativas coordenadas pelo Ministério do Meio Ambiente (MMA) e limitadas somente a dois anos – 2002 e 2013, respectivamente. Esses programas trabalham com mapeamentos envolvendo etapas e processos dependentes de máquinas de alto desempenho e de analistas (intérpretes) em todas as fases de construção desses produtos.

Atualmente, o advento de tecnologias de processamento de grande volume de dados, representados pelo conceito de *big data* e o processamento na nuvem (Alencar et al. 2020; Hashem et al. 2015; Shelestov et al. 2017) implicam na reformulação dessas metodologias “estáticas” de mapeamentos de uso e cobertura de terras. Essa estratégia de trabalho envolve o acesso e o uso de uma grande quantidade de dados primários e secundários e depende do quanto essas informações estão organizadas,

disponíveis e processáveis, o que implica, necessariamente, na existência de uma infraestrutura física e virtual capaz de suprir as demandas do usuário final, de acordo com suas necessidades (Yang et al. 2017). Neste sentido, a computação na nuvem vem se apresentando como solução na área de geoprocessamento ao possibilitar não apenas o armazenamento e acesso à grande quantidade de informação, mas também sua gestão e processamento em análises complexas com grande economia de tempo e recursos humanos e financeiros (Hashem et al. 2015; Yang, Xu & Nebert 2013; Yang et al. 2017).

Contudo, o uso de *big data* e computação na nuvem nas geociências apresenta também seus desafios. A curva de aprendizado por parte dos analistas de geoprocessamento bem como a implementação e operação da infraestrutura necessária para as análises – tais como servidores, redes, armazenamento, serviços e aplicações – são barreiras a serem superadas, levando pesquisadores a uma busca por soluções (Yang et al. 2017).

Mais recentemente, o Projeto de Mapeamento Anual da Cobertura e Uso do Solo no Brasil (MapBiomias), uma iniciativa em rede colaborativa formada por organizações não governamentais, universidades públicas, institutos de pesquisa e empresas privadas, deu início ao mapeamento dos biomas brasileiros em escala nacional, produzindo mapeamentos anuais de uso e cobertura do solo desde 1985, a partir de imagens da série histórica do satélite Landsat que são processadas pelas técnicas de *machine learning* na plataforma *Google Earth Engine* (GEE) (Alencar et al. 2020; Grande, Aguiar & Machado 2020; Souza et al. 2020; Wang et al. 2020).

Tanto as iniciativas em escala nacional quanto outros trabalhos em áreas menores identificaram algumas dificuldades para o mapeamento do uso e cobertura do solo no Cerrado (Alencar et al. 2020; Chelotti 2017; Nunes & Roig 2015; Sano et al. 2019). Tais dificuldades são resultantes das características climáticas e físicas específicas do bioma: forte sazonalidade climática, cobertura persistente de nuvens na estação chuvosa, transição gradual entre diferentes fitofisionomias e confusão de classes em função da dificuldade na diferenciação entre algumas formações naturais e áreas de atividades agropecuárias.

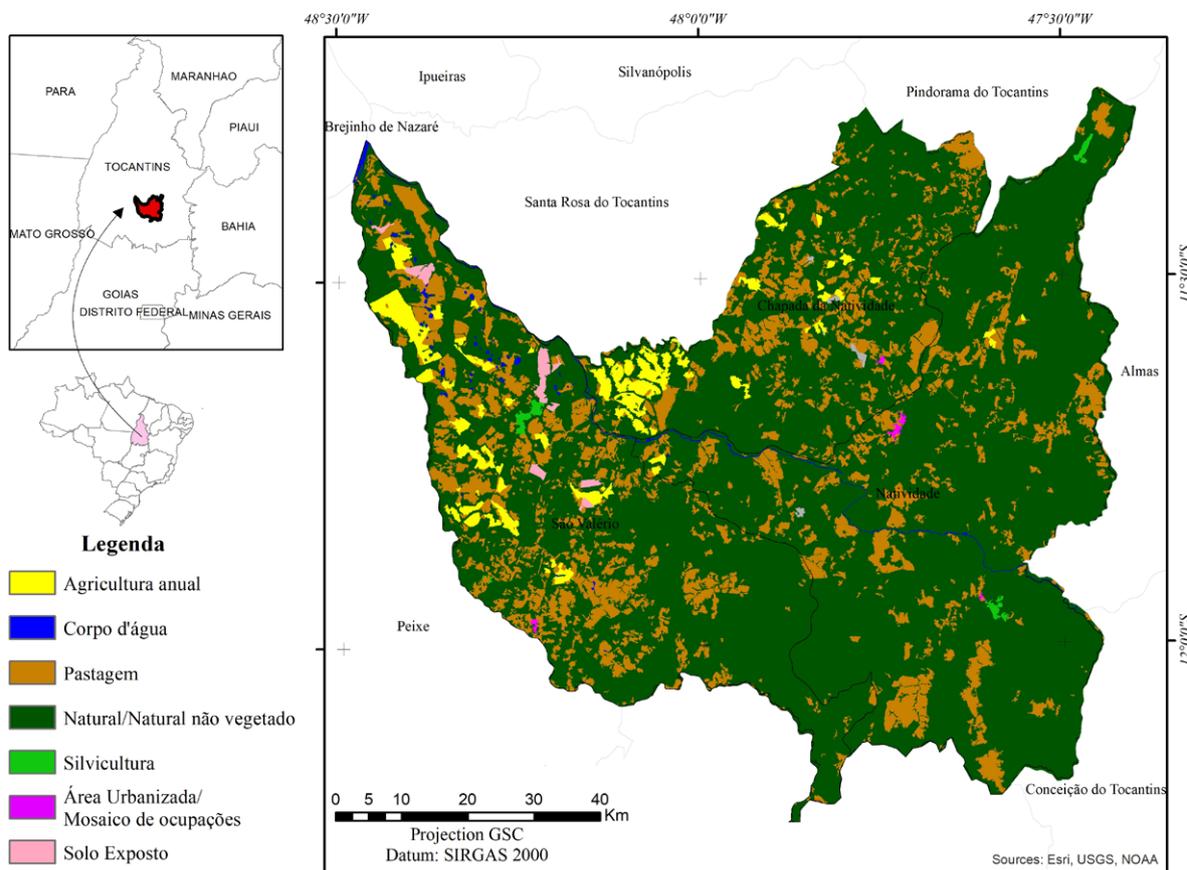
Considerando a relevância da produção de mapas acurados anuais do uso e cobertura do solo para o Cerrado, bem como do desenvolvimento de ferramentas e processos que demandem menos recursos humanos, financeiros e menor tempo para produção, o presente trabalho buscou avaliar o desempenho das ferramentas disponíveis na plataforma GEE e do classificador de imagens *Random Forest*, ambos utilizados nos processos de mapeamento de uso e cobertura do solo do projeto MapBiomass, por meio da simulação do mapeamento realizado pelo projeto TerraClass Cerrado para o ano de 2013 em uma região de Cerrado sob forte pressão antrópica. Este estudo foi baseado na integração de imagens adquiridas pelo satélite Landsat (dados multiespectrais), pelo sensor *Visible Infrared Imaging Radiometer Suite* (VIIRS) (luzes noturnas) dos satélites Suomi National Polar-Orbiting Partnership (Suomi NPP) e NOAA-20 e pela missão *Shuttle Radar Topography Mission* (SRTM) (modelo digital de elevação). A hipótese testada foi a de que a combinação de dados multiespectrais, de luzes noturnas e de elevação do terreno

apresenta desempenho superior na discriminação de classes de uso e cobertura do solo presentes na área de estudo em comparação aos métodos baseados apenas em informação espectral isolada.

## 2 Materiais e Métodos

### 2.1 Área de Estudo

A área de estudo situa-se integralmente no bioma Cerrado, mais especificamente, nos municípios de Natividade, Chapada da Natividade e São Valério da Natividade (estado de Tocantins), com aproximadamente 740.683 ha (Figura 1). Essa área faz parte de uma das 46 regiões definidas pelo Ministério do Meio Ambiente (MMA) como prioritárias para conservação de espécies ameaçadas de extinção, no âmbito do Projeto Estratégia Nacional para Conservação de Espécies Ameaçadas de Extinção - GEF-Pró-Espécies, com identificação de 29 espécies ameaçadas de extinção (Brasil 2016).



**Figura 1** Mapa de localização da área de estudo sobreposto ao mapa de uso e cobertura do solo do projeto TerraClass Cerrado (Brasil 2015).

Os municípios de Natividade, Chapada da Natividade e São Valério da Natividade, segundo os dados de uso e cobertura do solo do Projeto TerraClass Cerrado (Brasil 2015) possuem 78% da área coberta por vegetação natural, 18% por pastagem e 3% por agricultura anual. Não existem cadastros de unidades de conservação ou terras indígenas na região na base de dados do Ministério do Meio Ambiente (MMA) e da Fundação Nacional do Índio (FUNAI). A população estimada é de 16.485 habitantes com densidade demográfica média é de 2,17 hab/km<sup>2</sup> e Índice de Desenvolvimento Humano Municipal (IDHM) médio de 0,645 (Instituto Brasileiro de Geografia e Estatística 2020). Segundo dados de perfil econômico dos municípios de Tocantins (Secretaria do Planejamento e Orçamento 2017), os três municípios apresentam potencial de uso intensivo do solo para pecuária e cultivos agrícolas.

A região de estudo está incorporada na nova fronteira para expansão da agricultura no Brasil, denominada Matopiba (Maranhão, Tocantins, Sul do Piauí e Oeste da Bahia). Essa expansão está intimamente ligada ao desmatamento da vegetação nativa, pois os processos de regulamentação ambiental do Cerrado são menos rígidos quando comparados com os do bioma Amazônia (Spera et al. 2016).

Na área de estudo, os dados de precipitação (Figura 2; Instituto Nacional de Meteorologia 2020) para a estação de Santa Rosa, município fronteiriço, demonstram precipitação superior a 1.000 mm anuais com estações bem definidas em termos de períodos secos e chuvosos. A elevada temperatura média (Figura 2), que pode comprometer a oferta hídrica em plantações, pode também acelerar o crescimento vegetal, encurtando o ciclo produtivo. No entanto, a dinâmica da agricultura exige maior intensidade de capital e de conhecimento do que em regiões agrícolas

mais tradicionais do Brasil, como são os casos do Centro-Oeste e Sul (Garcia & Vieira Filho 2018).

Outro fator positivo para a produção rural na área de estudo é o fato de 90% da região possuírem declividade inferior a 8°, indicando impedimento nulo ou ligeiro à mecanização agrícola e muito baixo risco de erosão. A exceção é o município de Natividade, que possui 16% da sua área municipal com declividade acentuada.

## 2.2 Abordagem Metodológica

Todas as análises foram realizadas na plataforma GEE (Gorelick et al. 2017), voltada para análise de dados em escala planetária. A plataforma concentra um grande volume de dados, da ordem de Petabytes, provenientes de diferentes fontes como modelo digital de elevação da missão SRTM, dados históricos da série Landsat, *Moderate Resolution Imaging Spectroradiometer* (MODIS), Sentinel-1 e Sentinel-2, dentre outros catálogos de dados ou imagens.

A plataforma GEE também possibilita processamento de dados em nuvem de seus catálogos nativos, assim como a importação e exportação de dados matriciais e vetoriais. Trata-se de uma ferramenta de programação, cuja interação com os usuários é feita por meio de desenvolvimentos de *scripts* em linguagem de programação Javascript ou Python, assim como utilização de *scripts* já disponíveis ou com a utilização de *scripts* já disponíveis, i.e., desenvolvidos por outros usuários. Na plataforma, já foram implementadas diversas funções, desde operações matriciais simples até classificação de imagem orientada-a-objeto que vem sendo aplicada em trabalhos de mapeamento do uso e cobertura do solo (Alencar et al. 2020; Huang et al. 2017; Shelestov et al. 2017). Novas funções e novos catálogos de dados são constantemente incorporados. Novas implementações de inteligência artificial estão em desenvolvimento pela equipe de desenvolvimento (Google 2019).

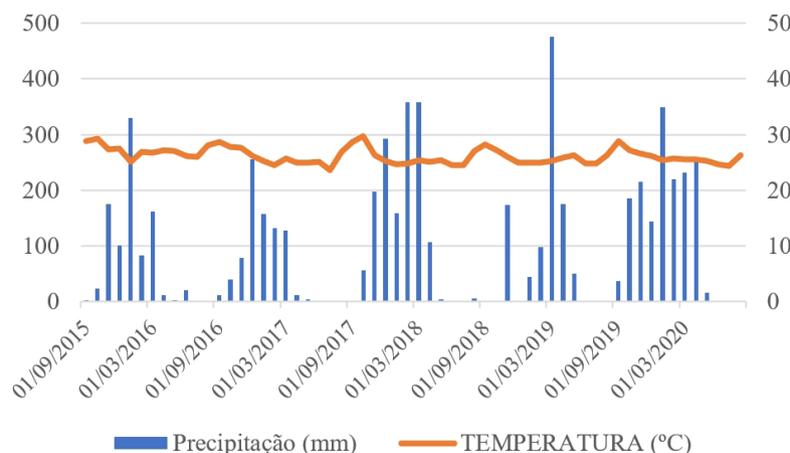


Figura 2 Dados de precipitação (barras em azuis) e de temperatura (linha vermelha) da estação de Santa Rosa do Tocantins (TO).

Conforme mostrada na Figura 3, o trabalho utilizou três dados de entrada já disponíveis na plataforma GEE: imagens multiespectrais do satélite Landsat 8 *Operational Land Imager* (OLI), imagens de luzes noturnas obtidas pelo sensor *Visible Infrared Imaging Radiometer Suite* (VIIRS) a bordo dos satélites *Suomi National Polar-Orbiting Partnership* (Suomi NPP) e NOAA-20, e o modelo digital de elevação obtido pela missão SRTM. Os dados de Landsat 8 OLI (órbitas/pontos 221/068, 221/069 e 222/068) foram obtidos entre maio e outubro de 2013, mesmo período do monitoramento do projeto TerraClass Cerrado 2013. Todos os registros de nuvem ou sombra de nuvem nas imagens Landsat foram eliminados por uma máscara baseada nos dados de qualidade radiométrica disponível no arquivo denominado 'pixel\_qa'.

Os dados de luzes noturnas foram utilizados como indicador de áreas urbanas (Sharma et al. 2016; Wang et al. 2017). Foi extraído o valor da mediana de todos os registros do ano a fim de reduzir possíveis ruídos derivados de queimadas, as quais ocorrem em maior quantidade no período da seca no Cerrado (maio a outubro). Os modelos digitais de elevação do SRTM foram utilizados para gerar mapas de elevação e declividade. A partir das imagens Landsat 8, foram extraídos os índices de água normalizados pela diferença (NDWI:  $(\rho\text{NIR} - \rho\text{SWIR})/(\rho\text{NIR} + \rho\text{SWIR})$ ) para extração dos corpos d'água (Mcfeeters 1996) e os índices de vegetação normalizados pela diferença (NDVI:  $(\rho\text{NIR} - \rho\text{RED} / \rho\text{NIR} + \rho\text{RED})$ ) para detecção de tipos de vegetação (Gao 1996). Os seguintes atributos foram

extraídos das imagens Landsat 8: a) NDWI: mediana; b) NDVI: mediana, moda, mínimo e máximo; e c) bandas espectrais: mediana, moda, mínimo, máximo e reflectância acumulada. A partir desses atributos, foi produzida uma imagem anual que corresponde a um cubo de imagem com 38 bandas (camadas) (Tabela 1).

O cubo de imagem anual foi então submetido à classificação orientada-a-objeto utilizando-se do classificador *Random Forest* (Breiman 2001). Nesse classificador, baseado no princípio de árvores de decisão, sub-amostras selecionadas aleatoriamente com repetição são utilizadas como variáveis preditivas para treinamento do classificador. Um terço dessas sub-amostras é selecionado para a etapa de validação, conhecida como "out of the bag". Um variado número de árvores é construído na etapa de treinamento, combinando-as para se ter uma predição com maior acurácia e estabilidade. Cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores da floresta. Durante a classificação, cada árvore escolhe, para cada pixel, a classe temática com maior probabilidade de acerto; a classe de uso e cobertura de solo mais votada é retornada pelo classificador (Belgiu & Drăguț 2016; Han, Kamber & Pei 2011). Belgiu & Drăguț (2016) apresentaram as principais aplicações do *Random Forest* em sensoriamento remoto, evidenciando as práticas realizadas e elencando pontos de atenção para seu uso, como a necessidade de uso de aproximadamente 0,25% de pontos amostrais para treinamento do classificador.

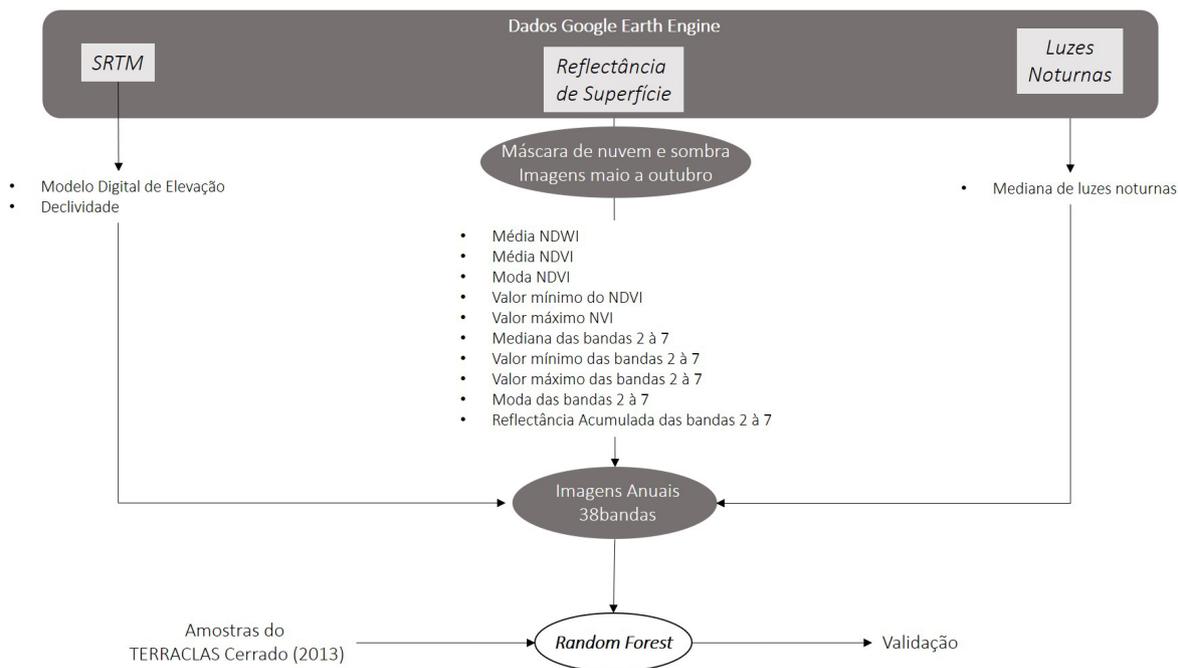


Figura 3 Fluxograma de processamento dos dados na plataforma Google Earth Engine.

Neste estudo, o classificador *Random Forest* foi executado a partir de 500 árvores de decisão, conforme sugerido por Belgiu & Drăguț (2016). Para o treinamento do algoritmo, foram utilizados pontos aleatórios para cada classe do mapeamento de uso e cobertura do solo do projeto TerraClass Cerrado 2013, previamente importado para o GEE, respeitando os limites de 30 m de cada pixel do Landsat, menor unidade amostral da classificação. Foram consideradas nove classes temáticas para a área de estudo. As classes ‘não observado’ e ‘mineração’ do projeto TerraClass Cerrado não foram consideradas devido à sua baixa representatividade espacial na área de estudo (Tabela 2). As classes que não atingiram 0,25% foram em decorrência da restrição da plataforma GEE ao usuário padrão. Desta forma, as classes predominantes, ‘natural’ e ‘pastagem’, foram subamostradas em relação à sugestão de Belgiu & Drăguț (2016).

O resultado da classificação foi submetido a uma interpolação modal considerando oito pixels adjacentes, por se tratar de um dado categórico, a fim de homogeneizar as classificações de pixels isolados (Deines et al. 2019). A análise da acurácia da classificação foi realizada por meio de matriz de confusão, acurácia global e índice Kappa. A amostragem foi do tipo estratificada não-alinhada, baseada

em um total de 1.000 pontos amostrais e no mapeamento do projeto TerraClass Cerrado 2013, o qual foi considerado como verdade terrestre (Figura 2).

### 3 Resultados e Discussão

Apesar dos atrativos para as atividades agrícolas tais como temperatura, pluviosidade e declividade favoráveis, a classificação resultante da metodologia proposta na área de estudo (Figura 4) demonstra evidente predominância de paisagem natural (94% para Natividade; 86% para Chapada da Natividade e 79% para São Valério da Natividade), com concentração de classes antrópicas somente na região noroeste dos municípios de São Valério da Natividade e Chapada da Natividade.

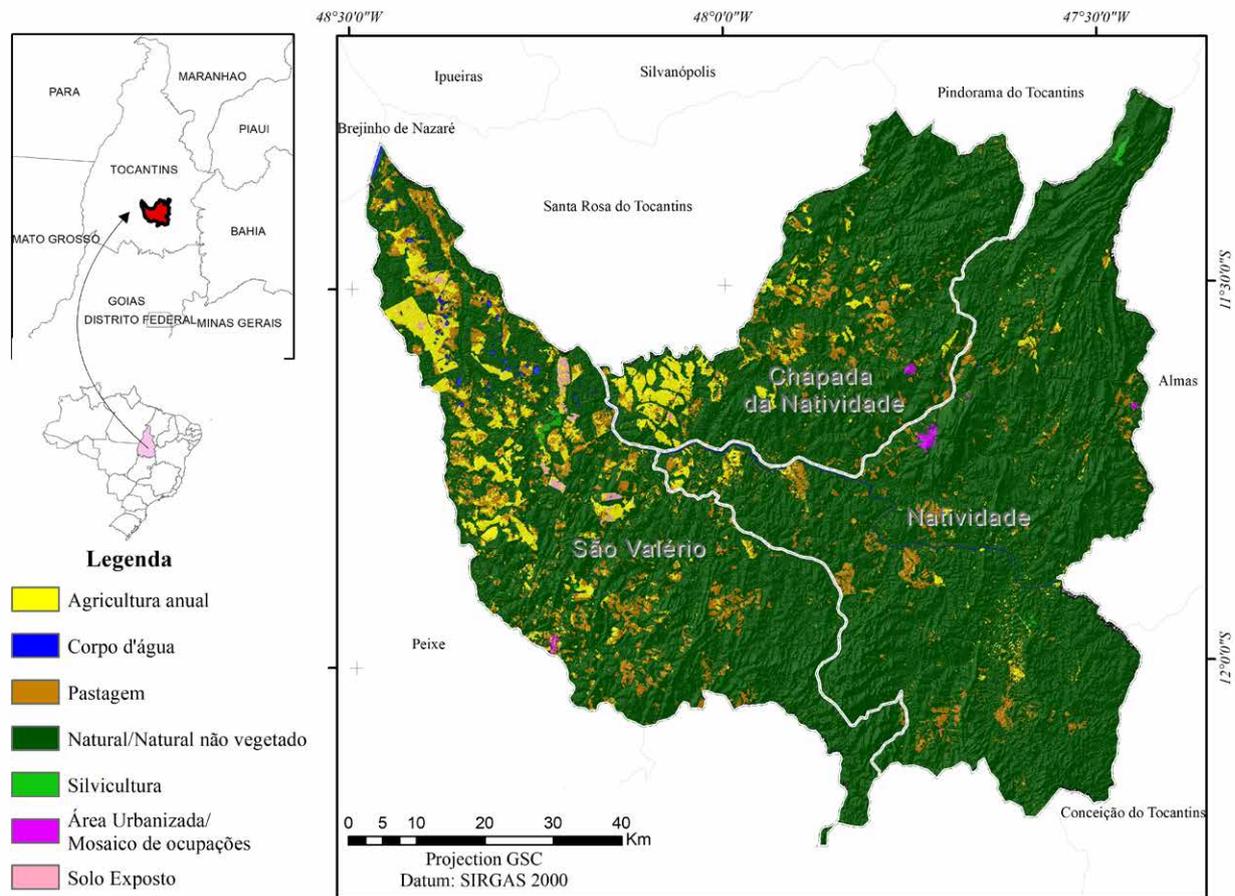
Tal resultado é similar ao visualizado nos resultados do MapBiomas – coleção 4.1. As coleções desse programa possuem tendência crescente de atividades antrópicas agropastoris entre os anos de 1985 e 2018, principalmente no município de São Valério da Natividade, em direção ao município da Chapada da Natividade. Na classificação para 2013 (Figura 4), ainda se observou uma baixa ocupação ao sul do município de Natividade, porém, houve um padrão de aumento para os anos subsequentes, de acordo com o mapeamento temporal do MapBiomas (MapBiomas 2019).

**Tabela 1** Dados de entrada do cubo de imagens com 38 bandas que foram utilizados para compor a imagem anual. As imagens do satélite Landsat-8 foram adquiridas no período de maio a outubro de 2013 (total de 12 imagens). SRTM = *Shuttle Radar Topography Mission*; VIIRS = *Visible Infrared Imaging Radiometer Suite*.

Fonte	Descrição	Resolução espacial	Dados primários ou derivados
Landsat- 8 OLI	Reflectância de superfície	30 m	NDWI (mediana) NDVI (mediana, moda, mínimo e máximo) Bandas espectrais (B2 a B7) (mediana, moda, mínimo, máximo e reflectância acumulada)
SRTM	Modelo digital de elevação	30 m	Elevação Declividade
VIIRS	Radiância média mensal	450 m	Mediana da série anual

**Tabela 2** Amostras de treinamento para a classificação de cubo de imagens da área de estudo.

Classe	Amostras de treinamento	Número de pixels	% de pontos amostrais para treinamento do classificador
Agricultura anual	600	243611	0,24
Corpos d'água	80	31065	0,25
Pastagem	800	1488907	0,053
Natural/Natural não-vegetado	4500	6622758	0,068
Silvicultura	80	19987	0,40
Área urbana/mosaico de ocupações	80	7130	1,12
Solo exposto	120	36389	0,33
Total	6260	8449847	2.461



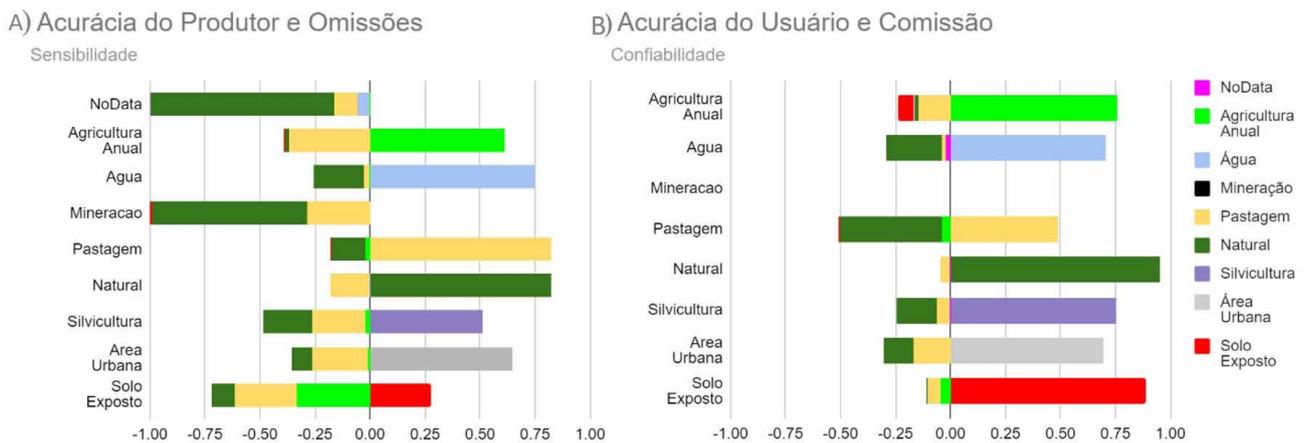
**Figura 4** Mapa de uso e cobertura do solo dos municípios de Natividade, Chapada da Natividade e São Valério da Natividade no estado de Tocantins para o ano de 2013, produzido a partir da classificação de cubo de 38 imagens pelo algoritmo *Random Forest* disponível na plataforma *Google Earth Engine*.

Tais evidências mostram que a pressão de atividades agrossilvopastoris na região tem aumentado nos últimos anos, provavelmente relacionados com o incentivo atual do governo federal. Dados da Embrapa confirmam o crescimento no estado de Tocantins em taxas superiores a 25% ao ano nos últimos 4 anos (Campos et al. 2019).

A acurácia global da classificação proposta foi de 0,84. Já o índice Kappa atingido foi de 0,64, podendo ser categorizado como “muito boa”, de acordo com Landis e Koch (1977). A classe com melhor desempenho na classificação foi “Natural”, com 95% de correspondência com a classificação de referência do TerraClass, seguida de “Solo Exposto” (89%) e “Agricultura Anual” (76%). “Área Urbana” apresentou 70% de correspondência com o TerraClass, sendo que 17% e 13% da confusão detectada foram com as classes “Pastagem” e “Natural”, respectivamente. A classe “Água” teve 71% de correspondência com o TerraClass e 25% de confusão

com a classe “Natural”. A classe “Silvicultura” obteve 75% de correspondência com o TerraClass e confusão majoritária com a classe “Natural” (18%). A classe com menor desempenho foi “Pastagem”, com 49% de correspondência com o TerraClass. Ainda a classe pastagem se destacou por apresentar 46% de confusão com a classe “Natural”, que, nessa região, é predominantemente coberta por formações savânicas.

Os resultados podem ser visualizados na Figura 5, na qual a acurácia do usuário (b) está associada ao erro de comissão, que é o erro cometido ao atribuir um pixel à uma classe quando este pertence a alguma outra classe, referindo-se a uma delimitação excessiva da categoria. A acurácia do produtor (a) está associada ao erro de omissão, que ocorre quando deixamos de mapear um pixel da classe corretamente. As barras localizadas à direita do gráfico mostram as taxas de acerto e ao lado esquerdo observa-se os percentuais de classes com confusão.



**Figura 5** A. Acurácia do produtor ou erro de omissão das classes temáticas consideradas neste estudo; B. acurácia do usuário ou erro de comissão.

Para a classe “Agricultura Anual”, é importante frisar que foi utilizada a quantidade de unidades amostrais próximas à sugerida por Belgiu & Drăguț (2016). Mesmo assim, observou-se uma subestimação da área de agricultura anual, nos quais pixels pertencentes a essa classe foram incorporados erroneamente às classes pastagem, vegetação natural e solo exposto.

Foi observado um baixo desempenho na classificação da classe “Pastagem”. Esse comportamento também foi observado por outros trabalhos em que a confusão espectral ocorre principalmente com as formações campestres naturais (Alencar et al. 2020; Grande, Almeida & Cicerelli 2016; Nascimento & Sano 2010). Apesar de a classe “Silvicultura” apresentar baixa representatividade, com 1.751 ha em mais de 7.000 km<sup>2</sup>, a classificação foi satisfatória, com 75% de correspondência com a classificação do projeto TerraClass Cerrado.

Conforme observado na Figura 5A, houve erros de omissão associados à classe solo exposto, principalmente com pastagem e agricultura anual. Tal ocorrência pode estar relacionada com a mudança natural dessa classe ao longo do ano, principalmente com agricultura anual. Essa constatação vai ao encontro da classe agricultura anual, cuja omissão ocorreu principalmente na classe pastagem.

É importante observar que tanto as amostras de treinamento quanto às amostras de validação foram obtidas a partir de pontos aleatórios distintos em ambos os processos, porém, advindos dos dados do TerraClass Cerrado. Assim, é possível que haja influência de eventuais erros de classificação nos dados considerados como referência, seja por erro de interpretação ou decorrente da área mínima utilizada nessa classificação de 6,25 ha. Certas inconsistências podem também estar associadas a mudanças temporais naturais dos pixels ao longo do ano,

uma vez que o mapeamento do projeto TerraClass Cerrado foi baseado em uma única cena por órbita/ponto. Uma forma de minimizar essas inconsistências seria por meio de uma análise temporal dessas classes.

Tendo em vista que os pontos gerados pelo *Random Forest* são aleatórios e as respostas espectrais são altamente influenciadas pela sazonalidade e variações anuais e regionais, os valores e regras de decisão no treinamento para o mapeamento de 2013 não foram aplicados para outros anos ou regiões. Ainda que alguns estudos tenham aplicado as regras de decisão para outros anos (Guerra, Schultz & Sanches 2017; Parente & Ferreira 2018), tais trabalhos focaram no mapeamento de uma única classe, como pastagem ou cultura anual, ao passo que o presente trabalho utilizou sete classes para classificação, aumentando, portanto, a probabilidade de erros de omissão e comissão. No caso de mapeamentos de vegetação natural ou mapeamentos envolvendo várias classes, entende-se que deve haver o treinamento do classificador em regiões de controle validadas em visitas de campo ou imagens de alta resolução, correspondentes ao ano de mapeamento (Alencar et al. 2020).

O bom desempenho do *Random Forest* na simulação do mapeamento do TerraClass reforça seu potencial para classificação do uso e cobertura do solo, em especial para a vegetação natural. Embora o desempenho do classificador não tenha sido superior, mas similar a outros métodos de classificação já aplicados no Cerrado (Chelotti 2017; Grande, Almeida & Cicerelli 2016; Nunes & Roig 2015), a metodologia aqui apresentada pode ser aperfeiçoada ao se adicionar outras variáveis espectrais e um treinamento de classificador mais preciso. Alencar et al. (2020), no âmbito do projeto MapBiomias, mapearam o uso e cobertura do solo em todo o bioma Cerrado nos anos de 1985 a 2018, também

a partir de imagens Landsat, fazendo uso da plataforma GEE e do classificador Random Forest, resultando em mapas com acurácia variado de 71% a 87%. O processo, no entanto, foi mais complexo e dispendioso por envolver estabelecimento de pontos de controle em campo e construção de árvores de decisão empírica e estatística, bem como processo de *machine learning*.

Ainda que a interface e operacionalização da plataforma exijam conhecimento de programação linear, por disponibilizar dados e permitir processamento na nuvem, as análises dispensaram a aquisição de dados, tornando o processo menos dispendioso em termos de tempo de preparação, processamento e análise. Para este estudo, foi possível extrair informações de cerca de 18 GBytes de imagens, sem contabilizar os dados do sensor VIIRS e da missão SRTM. A execução de todas as etapas, a partir de um *script* pronto, consumiu não mais que uma hora, sendo que o *script* pode ser adaptado para outras regiões, exigindo-se apenas a redefinição da área de estudo. A título de comparação, o TerraClass Cerrado foi desenvolvido ao longo de um ano (2014 a 2015) com várias instituições e equipes de trabalho, bem como equipamentos e *softwares*, utilizando apenas 121 cenas Landsat (visão estática) para todo o Cerrado (Brasil 2015). Neste estudo, foram utilizadas informações de 31 cenas Landsat 8 para uma região que corresponde a menos de 0,5% do bioma.

Recomenda-se que, para trabalhos futuros, seja utilizado outro método de coleta de elementos amostrais para treinamento do classificador. Assim, reforça-se a necessidade de reconhecimento da área de estudo em campo. Outras combinações de variáveis e dados podem ser testadas, por exemplo, o uso de imagens Sentinel e variáveis multiespectrais e espaciais distintas. Além do aprimoramento do classificador no bioma Cerrado, deve-se testar o método em outros biomas, já que as fitofisionomias apresentam especificidades que devem ser avaliadas e incorporadas no método. Considerando-se que há outros tipos de classificadores na plataforma, é recomendável testar outros algoritmos disponíveis.

## 4 Conclusões

Dentre os municípios mapeados, São Valério apresentou maiores áreas de classes antrópicas de uso e cobertura do solo para o ano de 2013 na região nordeste do município. Assim, é importante que o município monitore e fiscalize a ocupação do solo, buscando o cumprimento da legislação ambiental nos próximos anos.

Embora a proposta da pesquisa tenha buscado combinar dados multiespectrais com outros dados espaciais para aprimoramento da qualidade do mapeamento do uso

e cobertura do solo na região, os resultados se mostraram incipientes, pois a qualidade obtida foi similar à de pesquisas que utilizam apenas dados multiespectrais. Apesar disso, a metodologia mostrou-se vantajosa, pois processos de seleção de imagens, preparação, geração de *scripts* e processamento de dados são relativamente rápidos quando comparados com os procedimentos convencionais. Após a elaboração do *script* a ferramenta não consumiu mais que uma hora de processamento. Além do mais, deve-se ressaltar a ausência de custos financeiros para sua execução.

O resultado da classificação foi considerado satisfatório e sua acurácia se mostrou sensível à quantidade de amostras de treinamento e às variáveis a serem consideradas para o treinamento. Sugere-se a realização de testes adicionais com maior controle na obtenção dos parâmetros de entrada e de validação do classificador. Isto porque uma das possíveis razões desse desempenho do classificador foi o uso da classificação do TerraClass do ano de 2013 para treinamento e validação, o que pode ter influenciado na qualidade dos resultados.

Os resultados aqui apresentados reforçam o potencial do *Random Forest* para o mapeamento do bioma Cerrado. Com isso, apresentamos uma alternativa de baixo custo para mapear áreas menores no bioma Cerrado onde não há mapeamento atualizado disponível.

## 5 Referências

- Alencar, A., Shimbo, J.Z., Lenti, F., Marques, C.B., Zimbres, B., Rosa, M., Arruda, V., Castro, I., Ribeiro, J.P.F.M., Varela, V., Alencar, I., Piontekowski, V., Ribeiro, V., Bustamante, M.M.C., Sano, E.E. & Barroso, M. 2020, 'Mapping three decades of changes in the Brazilian savanna native vegetation using Landsat data processed in the Google Earth Engine platform', *Remote Sensing*, vol. 12, no. 6, 924. <https://doi.org/10.3390/rs12060924>
- Belgiu, M. & Drăgut, L. 2016, 'Random forest in remote sensing: A review of applications and future directions', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Brasil 2015, *Mapeamento do Uso e Cobertura do Cerrado: Projeto TerraClass Cerrado 2013*, Ministério do Meio Ambiente, Brasília, DF, 67 p.
- Brasil 2016, *Projeto GEF Pró-espécies*, viewed 12 August 2019, <<https://antigo.mma.gov.br/biodiversidade/economia-dos-ecossistemas-e-da-biodiversidade/item/11642-projeto-gef-pro-esppecies.html>>.
- Breiman, L. 2001, 'Random Forests', *Machine Learning*, vol. 45, pp. 5–32. <https://doi.org/10.1023/A:1010933404324>
- Campos, L.J.M., Costa, R.V., Almeida, R.E.M., Evangelista, B.A., Simon, J., Silva, K.J.N., Pereira, A.A. & Evaristo, A.B. 2019, *Produtividade de cultivares de soja em três ambientes do Tocantins*. Embrapa Soja, Londrina, PR, 18 p. (Boletim de Pesquisa e Desenvolvimento, 21).

- Chelotti, G.B. 2017, 'Mapeamento de uso do solo da bacia hidrográfica do Alto Descoberto, no Distrito Federal, por meio de classificação orientada a objetos com base em imagem do satélite Landsat 8 e softwares livres', *Revista Brasileira de Geomática*, vol. 5, no. 2, pp. 172–185. <http://dx.doi.org/10.3895/rbgeo.v5n2.5417>
- Deines, J.M., Kendall, A.D., Crowley, M.A., Rapp, J., Cardille, J.A. & Hyndman, D.W. 2019, 'Mapping three decades of annual irrigation across the US High Plains Aquifer using Landsat and Google Earth Engine', *Remote Sensing of Environment*, vol. 233, 111400. <https://doi.org/10.1016/j.rse.2019.111400>
- Gao, B.C. 1996, 'NDWI – A normalized difference water index for remote sensing of vegetation liquid water from space', *Remote Sensing of Environment*, vol. 58, no. 3, pp. 257–266. [https://doi.org/10.1016/S0034-4257\(96\)00067-3](https://doi.org/10.1016/S0034-4257(96)00067-3)
- Garcia, J.R. & Vieira Filho, J.E.R. 2018, 'O papel da dimensão ambiental na ocupação do MATOPIBA', *Confins*, no. 35. <https://doi.org/10.4000/confins.13045>
- Google 2019, *Google Earth Engine User Summit 2018*, viewed 3 June 2019, <<https://sites.google.com/earthoutreach.org/ceus2018/home?authuser=0>>.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. 2017, 'Google Earth Engine: Planetary-scale geospatial analysis for everyone', *Remote Sensing of Environment*, vol. 202, pp. 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Grande, T.O., Aguiar, L.M.S. & Machado, R.B. 2020, 'Heating a biodiversity hotspot: connectivity is more important than remaining habitat', *Landscape Ecology*, vol. 35, pp. 639–657. <https://doi.org/10.1007/s10980-020-00968-z>
- Grande, T.O., Almeida, T. & Cicerelli, R.E. 2016, 'Classificação orientada a objeto em associação às ferramentas reflectância acumulada e mineração de dados', *Pesquisa Agropecuária Brasileira*, vol. 51, no. 12, pp. 1983–1991. <https://doi.org/10.1590/S0100-204X2016001200009>
- Guerra, J.B., Schultz, B. & Sanches, I.D. 2017, 'Mapeamento automático da expansão da agricultura anual no MATOPIBA entre 2002 e 2015 utilizando a plataforma Google Earth Engine', *XVIII Simpósio Brasileiro de Sensoriamento Remoto*, Santos, SP, INPE, pp. 6850–7.
- Han, J., Kamber, M. & Pei, J. 2011, *Data Mining: Concepts and techniques*, 3<sup>rd</sup> ed., Morgan Kaufmann Publishers, Waltham, MA, USA.
- Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. & Khan, S.U. 2015, 'The rise of "big data" on cloud computing: Review and open research issues', *Information Systems*, vol. 47, pp. 98–115. <https://doi.org/10.1016/j.is.2014.07.006>
- Huang, H., Chen, Y., Clinton, N., Wang, J., Wang, X., Liu, C., Gong, P., Yang, J., Bai, Y., Zheng, Y. & Zhu, Z. 2017, 'Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine', *Remote Sensing of Environment*, vol. 202, pp. 166–176. <https://doi.org/10.1016/j.rse.2017.02.021>
- Instituto Brasileiro de Geografia e Estatística 2020, *Conheça cidade e estados do Brasil*, viewed 29 August 2020, <<https://cidades.ibge.gov.br/>>.
- Instituto Nacional de Meteorologia 2020, *BDMEP - Dados históricos*, viewed 15 January 2021, <<https://portal.inmet.gov.br/servicos/bdmep-dados-hist%25C3%25B3ricos/>>.
- Klink, C. & Machado, R. 2005, 'A conservação do Cerrado brasileiro', *Megadiversidade*, vol. 1, no. 2, pp. 147–155.
- Landis, J.R. & Koch, G.G. 1977, 'An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers', *Biometrics*, vol. 33, no. 2, pp. 363–374. <https://doi.org/10.2307/2529786>
- MapBiomas 2019, *Coleção 4.1 da série anual de mapas de cobertura e uso de solo do Brasil*, viewed 13 November 2020, <[https://mapbiomas.org/colecoes-mapbiomas-1?cama\\_set\\_language=pt-BR](https://mapbiomas.org/colecoes-mapbiomas-1?cama_set_language=pt-BR)>.
- McFeeters, S.K. 1996, 'The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features', *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432. <https://doi.org/10.1080/01431169608948714>
- Nascimento, E.R.P. & Sano, E.E. 2010, 'Identificação de Cerrado Rupestre por meio de imagens multitemporais do Landsat: proposta metodológica', *Sociedade & Natureza*, vol. 22, no. 1, pp. 93–106.
- Nunes, J.F. & Roig, H.L. 2015, 'Análise e mapeamento do uso e ocupação do solo da bacia do alto do descoberto, DF/GO, por meio de classificação automática baseada em regras e lógica nebulosa', *Revista Arvore*, vol. 39, no. 1, pp. 25–36. <https://doi.org/10.1590/0100-67622015000100003>
- Parente, L. & Ferreira, L. 2018, 'Assessing the spatial and occupation dynamics of the Brazilian pasturelands based on the automated classification of MODIS images from 2000 to 2016', *Remote Sensing*, vol. 10, no. 4, 606. <https://doi.org/10.3390/rs10040606>
- Sano, E.E., Ferreira, L.G., Asner, G.P. & Steinke, E.T. 2007, 'Spatial and temporal probabilities of obtaining cloud-free Landsat images over the Brazilian tropical savanna', *International Journal of Remote Sensing*, vol. 28, no. 12, pp. 2739–2752. <https://doi.org/10.1080/01431160600981517>
- Sano, E.E., Rosa, R., Scaramuzza, C.A.M., Adami, M., Bolfe, E.L., Coutinho, A.C., Esquerdo, J.C.D.M., Maurano, L.E.P., Narvaes, I.S., Oliveira Filho, F.J.B., Silva, E.B., Victoria, D.C., Ferreira, L.G., Brito, J.L.S., Bayma, A.P., Oliveira, G.H. & Bayma-Silva, G. 2019, 'Land use dynamics in the Brazilian Cerrado in the period from 2002 to 2013', *Pesquisa Agropecuária Brasileira*, vol. 54, e00138. 10.1590/S1678-3921.pab2019.v54.00138
- Secretaria do Planejamento e Orçamento do Governo do Tocantins 2017, *Perfil socioeconômico dos municípios*, viewed 11 June 2019, <<http://www.sefaz.to.gov.br/estatistica/estatistica/indicadores-socioeconomicos/estatistica/indicadores-socioeconomicos/versao-2017/>>.
- Sharma, R.C., Tateishi, R., Hara, K., Gharechelou, S. & Iizuka, K. 2016, 'Global mapping of urban built-up areas of year 2014 by combining MODIS multispectral data with VIIRS nighttime light data', *International Journal of Digital Earth*, vol. 9, no. 10, pp. 1004–1020. <https://doi.org/10.1080/17538947.2016.1168879>
- Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A. & Skakun, S. 2017, 'Exploring Google Earth Engine platform for big

- data processing: Classification of multi-temporal satellite imagery for crop mapping', *Frontiers in Earth Science*, vol. 5. <https://doi.org/10.3389/feart.2017.00017>
- Souza, C.M., Shimbo, J.Z., Rosa, M.R., Parente, L.L., Alencar, A.A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., Ferreira, L.G., Souza-Filho, P.W.M., Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Ferreira Neto, L.C., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V. & Azevedo, T. 2020, 'Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat archive and Earth Engine', *Remote Sensing*, vol. 12, no. 17, 2735. <https://doi.org/10.3390/rs12172735>
- Spera, S.A., Galford, G.L., Coe, M.T., Macedo, M.N. & Mustard, J.F. 2016, 'Land-use change affects water recycling in Brazil's last agricultural frontier', *Global Change Biology*, vol. 22, no. 10, pp. 3405–3413. [10.1111/gcb.13298](https://doi.org/10.1111/gcb.13298)
- Vieira, R.R.S., Ribeiro, B.R., Resende, F.M., Brum, F.T., Machado, N., Sales, L.P., Macedo, L., Soares-Filho, B. & Loyola, R. 2018, 'Compliance to Brazil's Forest Code will not protect biodiversity and ecosystem services', *Diversity and Distributions*, vol. 24, no. 4, pp. 434–438. <https://doi.org/10.1111/ddi.12700>
- Wang, R., Wan, B., Guo, Q., Hu, M. & Zhou, S. 2017, 'Mapping regional urban extent using NPP-VIIRS DNB and MODIS NDVI data', *Remote Sensing*, vol. 9, no. 8, 862. <https://doi.org/10.3390/rs9080862>
- Wang, L., Diao, C., Xian, G., Yin, D., Lu, Y., Zou, S. & Erickson, T.A. 2020, 'A summary of the special issue on remote sensing of land change science with Google earth engine', *Remote Sensing of Environment*, vol. 248, 112002. <https://doi.org/10.1016/j.rse.2020.112002>
- Yang, C., Xu, Y. & Nebert, D. 2013, 'Redefining the possibility of digital Earth and geosciences with spatial cloud computing', *International Journal of Digital Earth*, vol. 6, no. 4, pp. 297–312. <https://doi.org/10.1080/17538947.2013.769783>
- Yang, C., Huang, Q., Li, Z., Liu, K. & Hu, F. 2017, 'Big data and cloud computing: Innovation opportunities and challenges', *International Journal of Digital Earth*, vol. 10, no. 1, pp. 13–53. <https://doi.org/10.1080/17538947.2016.1239771>

Recebido em: 10/09/2020

Aprovado em: 30/03/2021

### Como citar:

Cho, D.F., Schwaida, S.F., Cicerelli, R.E., Almeida, T., Ramos, A.P.M. & Sano, E.E. 2021. 'Desempenho do Algoritmo de Classificação de Imagens Random Forest para Mapeamento do Uso e Cobertura do Solo no Cerrado Brasileiro', *Anuário do Instituto de Geociências*, vol. 44: 37979. [https://doi.org/10.11137/1982-3908\\_2021\\_44\\_37979](https://doi.org/10.11137/1982-3908_2021_44_37979)