

DETECÇÃO DE ÁREAS DE FLORESTAS ESTÁVEIS EM SÉRIES TEMPORAIS UTILIZANDO RANDOM FOREST

Eduardo Ribeiro Lacerda^{1, 2}

Raúl Sánchez Vicens¹

1. Universidade Federal Fluminense - Departamento de Geografia - Campus da Praia Vermelha, Boa Viagem, Niterói, Brasil (eduardolacerdageo@gmail.com) (rsvicens@gmail.com)

2. Instituto Internacional para Sustentabilidade - Estrada Dona Castorina, Jardim Botânico, Rio de Janeiro, Brasil

ABSTRACT

The emergence of vegetation change detection algorithms in the last decade is impressive. But the results still have a lot of noise that needs to be cleaned. And the data cleaning process still uses other landcover mapping results. Besides that, the necessity to generate invariant land use classes is important to know particularly to forest areas. Thinking about that, this paper seeks to create a new form of mapping these invariant areas that can be used to mask noise and as an input on other conservation and restoration studies. The methodology proposed here uses the Google Earth Engine platform and a Random Forest classifier to classify invariant forest areas using all the images in the time series at once. The results found that the new approach found better results than the typical use of previous mapping. We found an overall accuracy of 91,7% using this method. Also, this paper seeks to contribute to the remote sensing community showing after exhaustive testing, good options of variables to use on this type of work.

Keywords: Séries Temporais, Detecção de Mudanças, Florestas, Google Earth Engine, Random Forest

INTRODUÇÃO

O monitoramento remoto da paisagem se estabeleceu em nossa sociedade como uma das principais formas de planejamento estratégico e como uma ferramenta importante para a quantificação de recursos e execução de políticas públicas. Os programas de monitoramento de queimadas e de desmatamento ilegal ganham cada vez mais espaço nos noticiários recentes não somente devido a clara relevância dos resultados obtidos, mas também porque seus resultados repercutem diretamente em muitos dos acordos comerciais que os países tem a possibilidade de realizar.

Dentre todas as possibilidades de uso de ferramentas de sensoriamento remoto para a detecção de mudanças da paisagem, a detecção de mudanças em áreas florestadas focando em eventos de supressão tem um papel importante e possivelmente é uma das mais estudadas e com maior foco na área. Além dos motivos apontados anteriormente, os dados de monitoramento das supressões ocorridas nessas áreas podem ser usados também como dado de entrada para análises que visam a priorização espacial de áreas estratégicas para a conservação e mais recentemente para a restauração de áreas com histórico de supressão. Projetos como o Mapbiomas (SOUZA et al., 2020) realizam o mapeamento sistemático de

uso do solo para todo o território nacional e a detecção de mudanças de uso ao longo do período do mapeamento (1985 – 2018), sendo um exemplo de como a área de geotecnologias pode se beneficiar com o aumento do acesso à computação de alto desempenho.

Além disso, devido a esse aumento do poder computacional e de seu acesso, novos algoritmos visando a análise de séries temporais utilizando imagens digitais orbitais como base surgiram como uma tentativa de aprimorar não só a detecção da mudança em si como de sua interpretação, agregando cada vez mais conteúdo de qualidade as análises. Alguns desses algoritmos visam a detecção de mudanças em áreas de floresta como o caso do CCDC - *Continuous Change Detection and Classification* (ZHU e WOODCOCK, 2014), COLD - *Continuous Monitoring of Land Disturbance* (COHEN et al., 2020), Landtrendr (KENNEDY et al., 2010; KENNEDY et al., 2012), VCT – *Vegetation Change Tracker* (HUANG et al., 2010; THOMAS et al., 2011), EWMACD - *Exponentially Weighted Moving Average Change Detection* (BROOKS et al., 2014), MIICA - *Multi-index Integrated Change Analysis* (JIN et al., 2013), VerDET - *Vegetation Regeneration and Disturbance Estimates through Time* (HUGHES et al., 2017), ITRA - *Image Trends from Regression Analysis* (VOGELMANN et al., 2012) e o Shapes-NBR (MEYER, 2013; MOISEN et al., 2016).

Pensando pela perspectiva temporal, é importante entender exatamente qual tipo de mudança se espera obter com a utilização desses algoritmos. Muitas das ferramentas já citadas possibilitam a detecção não somente de eventos de grande magnitude, mas também de processos de degradação que ocorreram ao longo de décadas, o que pode gerar resultados com muito ruído.

Como se pode imaginar, a limpeza dos dados na etapa de pós processamento desses algoritmos é essencial para a obtenção de resultados de boa qualidade. É normal que pequenas variações na própria radiometria da imagem sejam detectadas como mudança. Apesar de ser possível retirar grande parte desse ruído considerando apenas mudanças com magnitudes mais altas, áreas de floresta que possuem alta declividade, por exemplo, tendem a sofrer mais com certos ruídos. Sendo assim, parte da limpeza desses dados normalmente acontece utilizando dados de projetos como o Mapbiomas como base para ignorar áreas de não interesse, como áreas que foram classificadas como urbanas durante todos os anos do mapeamento. Ou então, com a retirada de todas as áreas que sempre foram classificadas como corpos d'água, pasto, solo exposto e claro, florestas. No entanto, os mapeamentos anuais do Mapbiomas entendem a classificação de forma anual, e não de forma multitemporal, o que pode implicar em erros quando a multiplicação das classificações é realizada para gerar as camadas binárias que servirão para limpar os ruídos.

Sendo assim, o presente trabalho busca elaborar uma alternativa a essa técnica, apresentando uma forma mais precisa de gerar camadas de áreas de estabilidade, ou seja,

áreas que apesar de terem sido afetadas por pequenas mudanças mantiveram alguma característica principal ao longo da série. Utilizamos para isso o algoritmo Random Forest na plataforma Google Earth Engine para realizar essa detecção não utilizando camadas classificadas ano a ano, mas classificando toda a série de uma só vez.

METODOLOGIA

Para a realização desse estudo utilizou-se como área de estudo o limite a APA da Bacia Hidrográfica do Rio São João. A área foi escolhida por já possuir extenso número de trabalhos de referência publicados e por possuir elementos da paisagem que representam de forma geral grande parte o contexto morfoclimático e fitogeográfico do Estado do Rio de Janeiro.

Inicialmente foi necessário gerar a camada de floresta invariante utilizando os dados do Mapbiomas. Ou seja, tudo que foi classificado como floresta durante todos os anos do mapeamento. Para isso, todas as camadas de uso do solo para o bioma da Mata Atlântica gerados pela série 4.1 do Mapbiomas foram recortados de acordo com os limites da APA e passaram por um processo de reclassificação resultando em uma camada binária (floresta e não-floresta). Após a reclassificação, todas as camadas foram multiplicadas entre si, resultando finalmente na camada binária final.

Como a proposta metodológica era a aplicação de um algoritmo de classificação em toda a série, a coleta de boas amostras do que representava as florestas invariantes era essencial. Além disso, não seria prudente utilizar somente a camada do Mapbiomas para coletar tais amostras, já que a mesma seria comparada com o resultado final. Sendo assim, além da utilização dos próprios dados do Mapbiomas, tivemos que gerar uma camada extra com o valor máximo de NDVI considerando todos os anos da série para aí sim fazer uma coleta de amostras através de uma abordagem mais conservadora.

A camada de máximo valor de NDVI foi toda desenvolvida utilizando a plataforma Google Earth Engine. Para a elaboração dessa camada foram utilizadas todas as imagens disponíveis do satélite Landsat considerando as séries 5, 7 e 8. Foram utilizadas todas as imagens disponíveis para todos os três satélites presentes na coleção Surface Reflectance Tier 1 por já possuírem correção atmosférica (6S).

Através da filtragem inicial uma função percorreu cada uma das imagens e gerou uma banda NDVI correspondente para cada uma delas. Além disso, uma outra função retirou todos os pixels com presença de nuvens e sobras.

Utilizando então a banda do NDVI foi possível agregar para cada pixel o valor da mediana considerando todas as imagens disponíveis para cada ano analisado, para aí sim, extrair os valores máximos considerando todos os anos da série. Ou seja, a camada final representa o valor máximo das medianas. A escolha da mediana como um passo intermediário é crucial para a eliminação de ruídos e valores inválidos presentes na série temporal. O cálculo

baseado apenas na aplicação da função de extração dos valores máximos acaba gerando uma camada final com valores ruidosos espalhados por toda a imagem, o que impossibilita seu uso.

Após a criação da camada de máximo valor do NDVI, foi necessário encontrar um limiar para o valor de corte do que representaria na camada um valor de floresta. Considerando que os tipos de uso do solo que mais podem gerar confusão em uma classificação com áreas de floresta são as áreas de pastagem e agricultura, uma nova camada contendo somente essas duas classes foi gerada utilizando os mesmos dados do Mapbiomas e a mesma metodologia que a realizada anteriormente com áreas de floresta.

Tanto a camada de florestas invariantes como a de pasto/agricultura invariantes foram transformados em pontos (vetor) e dois mil pontos para cada classe foram selecionadas de forma aleatória. Com isso, um teste t de Student foi realizado apresentando um valor-p de $2.2e-16$, mostrando uma diferença significativa entre as duas amostras. Após esta etapa, uma função para extrair o valor mínimo presente nas duas mil amostras de florestas foi feita apresentando um valor de 0,83. Sendo assim, uma camada (raster) final, igualmente binária considerando apenas valores maiores ou iguais a 0,83 foi criada. A multiplicação da camada binária de florestas invariantes do Mapbiomas com a camada binária de máximo NDVI serviu então como nossa camada para extração de pontos de floresta de treinamento para o modelo. Já para a extração de pontos de florestas que apresentaram alguma mudança significativa, utilizamos o resultado do algoritmo Landtrendr. A escolha do Landtrendr se deu devido a sua implementação mais madura na plataforma Google Earth Engine quando comparada a outras opções e devido a sua capacidade de detecção de eventos com baixa magnitude, o que em muitos casos significa a detecção de processos de degradação mais lentos. Dentre os muitos resultados que o algoritmo lhe oferece, a camada escolhida foi o da detecção dos maiores eventos de perda de vegetação (*Greatest Loss*) utilizando imagens anuais de 1985 até 2018 e considerando todas as imagens disponíveis para cada ano no período de 1 de janeiro até 31 de dezembro. Como camada padrão utilizou-se o NDVI para o processamento do resultado final.

Após o Landtrendr apresentar o resultado com todas as maiores perdas detectadas, iniciou-se o processo de limpeza desses dados ao considerar apenas mudanças com magnitudes maiores que 200, ou seja, perdas maiores que 0,2 no NDVI em um único evento. Nenhuma outra limitação foi imposta ao algoritmo e todos os outros parâmetros foram utilizados de forma padrão sem nenhuma modificação.

Já os pontos para a classe “outros” (água, solo exposto, pasto, etc.) foram coletados de forma aleatória utilizando o raster resultante com todos os pixels que não tinham sido classificados nem como floresta pela fusão da camada do Mapbiomas com o NDVI máximo, nem pelos pixels detectados pelo Landtrendr como áreas com possível perda de vegetação.

Todas as amostras das três classes foram então importadas em formato shapefile para a plataforma do Google para a então execução do processo de classificação utilizando o algoritmo Random Forest. Como imagem base para a classificação precisamos de uma série temporal com imagens anuais seguindo os mesmos padrões da série do Mapbiomas para possibilitar uma posterior comparação. Essa série de imagens foi gerada utilizando um *script* desenvolvido no Earth Engine contendo para cada ano uma composição de valores considerando o valor da mediana de cada pixel para cada banda. Ou seja, para cada ano tínhamos uma camada da banda Blue, Green, Red, NIR, SWIR1, SWIR2 e também de índices como o NDVI, NDWI, NDMI, SAVI, Greenness, Wetness e Brightness com valores da mediana considerando todos os pixels válidos de cada ano. Áreas com nuvens e sombra foram mascarados com valores “no data” e, portanto, não foram considerados no cálculo da mediana.

No entanto, por se tratar de uma nova metodologia, antes de gerar o processo de classificação final utilizando as amostras coletadas, escolhemos realizar testes de validação cruzada para entender melhor quais variáveis (bandas) teriam o melhor desempenho ao classificar este tipo de classe. Utilizamos então as amostras coletadas para extrair os valores da série temporal e exportamos para um ambiente *offline*, já que atualmente não existem ferramentas de teste e validação suficientemente boas no Earth Engine.

As tabelas com os valores das amostras foram processadas no pacote MLR (*Machine Learning in R*) (BISCHL, et al., 2016) utilizando o classificador Random Forest com 100 árvores cada e com a validação cruzada em modo “*k-fold*” utilizando parâmetro 10 (divide o subconjunto das amostras em 90% treino e 10% teste), repetindo ainda o processo por 100 iterações para cada teste, obtendo então um único valor de índice kappa para cada rodada. Cada rodada aqui representa uma combinação de bandas a serem utilizadas pelo classificador para tentar separar as três classes. Para este teste foram consideradas 24 combinações diferentes de bandas. Aplicando o Random Forest para as 24 variações de entrada com o processo de validação cruzada utilizando 10 “*folds*” e 100 iterações para cada validação, foram gerados 24000 processos de classificação diferentes, com cada processo gerando 100 árvores de decisão. Os resultados para essa quantidade exaustiva de testes são os apresentados na Tabela 1:

TABELA 1: ÍNDICE KAPPA PARA AS COMBINAÇÕES DE VARIÁVEIS UTILIZADAS

BANDA	Kappa
Blue + Green + Red + NIR + SWIR1 + SWIR2 + NDVI + NDWI + NDMI + SAVI	0,844754
Blue + Green + Red + NIR + SWIR1 + SWIR2 + NDVI + NDWI + NDMI + SAVI + Greenness + Wetness + Brightness	0,8398718
Blue + Green + Red + NIR + SWIR1 + SWIR2	0,8388044
Blue + Green + Red + NIR	0,8333813
Blue + Green + Red	0,8327737
Greenness + Wetness + Brightness + NDVI + NDWI + NDMI + SAVI	0,8295032
NDVI + NIR + NDMI	0,8198751
NDVI + NIR	0,8198327
Greenness + Wetness + Brightness	0,8126214
Red	0,8124956
SWIR2	0,8089151
Green	0,8055279
Blue	0,8000793
NDVI + NDWI + NDMI + SAVI	0,7966787
NDVI + NDMI	0,7833422
SWIR1	0,7788457
Wetness	0,7675543
Greenness	0,761319
NDVI	0,7554752
SAVI	0,7552823
Brightness	0,7159552
NDMI	0,7060223
NDWI	0,6343948
NIR	0,5760713

É importante frisar que o processo de validação cruzada representa apenas uma etapa intermediária antes da classificação final que visa uma melhor escolha dos parâmetros ao verificar através de muitos testes qual combinação de variáveis obteve melhor resultado (menor erro).

Como se pode observar, a melhor combinação de bandas para o processo de classificação de áreas de florestas estáveis, florestas que sofreram algum tipo de perda e outros não é a simples utilização de todas as bandas, apesar desta opção ter apresentado o segundo melhor

desempenho. A incorporação do Tasseled Cap não trouxe uma melhoria na acurácia do modelo. No entanto, a tabela de desempenho também apresenta outros resultados interessante, como por exemplo a taxa de acerto utilizando apenas a banda do vermelho (Red) ou em como bandas que são tipicamente usadas pela comunidade de forma ampla para detecção de áreas vegetadas como o NDVI apresentaram resultado pior do que muitas outras combinações possíveis.

RESULTADOS

Após a otimização do modelo, pode-se então fazer uma seleção (*subset*) na série temporal para utilizar apenas a combinação de bandas que obtiveram o melhor resultado (Blue, Green, Red, NIR, SWIR1, SWIR2, NDVI, NDMI, NDWI e SAVI) e assim gerar uma classificação final (Figura 1). Como resultado obtivemos uma área total de florestas invariantes na APA de 501,1km².

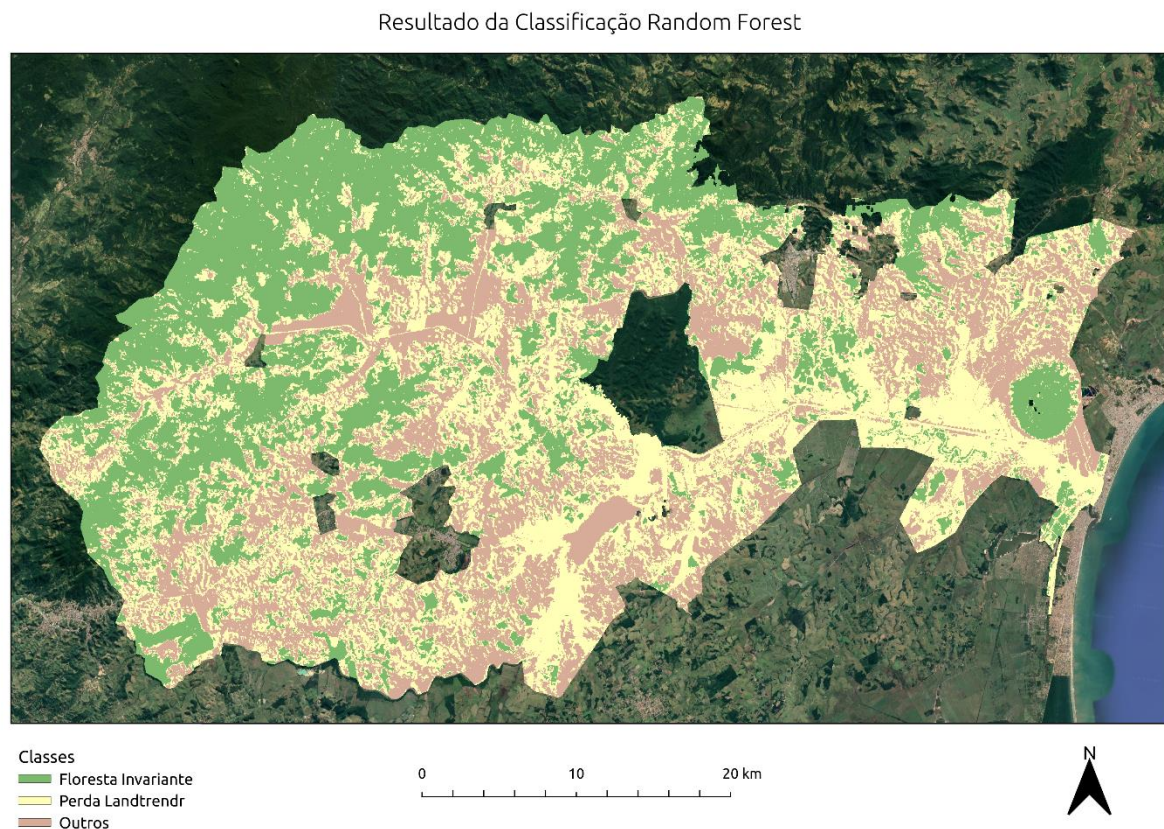


Figura 1. Classificação final gerada no Google Earth Engine (Classificação final)

Mas além da classificação gerada, uma nova imagem teve de ser criada para a comparação entre o resultado obtido pelo Random Forest e a antiga imagem binária utilizando somente dados do Mapbiomas. A nova imagem foi criada somando-se a camada de florestas estáveis com todos os anos do Mapbiomas com a versão binarizada do resultado final (florestas estáveis - outros).

Com essa imagem fusionada pode se notar que grande parte dos pixels (469,8km²) foram classificados como floresta invariante tanto pelo Random Forest como pelo Mapbiomas. No entanto, diferenças como áreas onde somente o Random Forest classificou como floresta (22,7km²) e onde somente o Mapbiomas classificou como floresta (52,7km²) apareceram. Para entender melhor a qualidade de cada resultado, foram feitos alguns testes estatísticos além da própria validação de cada resultado. Primeiramente, foram selecionadas duas mil amostras aleatórias de três classes: Áreas que foram classificadas como floresta pelos dois mapeamentos (FOR), áreas que foram classificadas como floresta apenas pelo Random Forest (RF) e áreas que foram classificadas como floresta apenas pelo Mapbiomas (MB). Um teste de análise de variância (ANOVA) foi realizado para buscar entender se os valores encontrados por cada amostra eram estatisticamente similares. O valor-p obtido pela ANOVA foi de 0,0116, o que mostrava uma diferença significativa entre as amostras considerando um limiar de valor-p de 0,05. Sendo assim, buscou-se então realizar um teste entre cada uma das amostras utilizando o teste de Tukey. Os resultados são os apresentados na Tabela 2:

TABELA 2: RESULTADOS DO TESTE DE TUKEY

Combinação (Teste de Tukey)	Valor-p
Mapbiomas - FOR	0,0008489
RF - FOR	0,1841643
RF - Mapbiomas	0,0000002

Os resultados foram promissores. Segundo o teste de Tukey, ao mesmo tempo em que as amostras que somente o Mapbiomas mapeou como floresta invariante são significativamente diferente das que ambos os mapeamentos mapearam como floresta invariante, e as amostras do Random Forest são diferente das do Mapbiomas, as amostras do Random Forest quando comparadas com as mapeadas por ambos são estatisticamente similares.

Este resultado mostra que a classificação feita pela metodologia proposta apresenta um resultado possivelmente superior a obtida utilizando somente as camadas do Mapbiomas. No entanto, ainda seria preciso saber o quanto de erro associado poderia existir não só nas áreas mapeadas pelo Mapbiomas como pelo Random Forest, mas como também na que ambos tiveram o mesmo resultado.

Para este processo de validação final utilizamos a ferramenta TimeSync (COHEN et al., 2010). Diferente do processo de validação comum nos mapeamentos de apenas uma ou poucas datas, a validação de séries temporais densas necessita de ferramentas especiais como o TimeSync, que apesar de ter uma utilização ainda consideravelmente complexa, é uma das únicas ferramentas existentes que realizam esse tipo de validação.

O TimeSync funciona de forma integrada com o Google Earth Engine e também necessita do software Access da Microsoft para armazenar as coordenadas de cada ponto a ser validado. A validação do TimeSync é feita de forma visual através da interface gráfica do programa (<https://github.com/eMapR/TimeSync-Legacy>).

Foram coletados para a etapa de validação no TimeSync 300 pontos para cada uma das classes. Os pixels mapeados apenas pelo Mapbiomas tiveram uma taxa de acerto de 55%, enquanto os pixels mapeados somente pelo Random Forest tiveram uma taxa de acerto de 76%. Já os pixels mapeados por ambos tiveram um acerto de 97%. Logo após, um último teste foi realizado considerando outras 300 amostras do mapa final gerado pela nova metodologia, mas considerando apenas os pixels que foram classificados como floresta invariante, resultado em uma taxa de acerto final de 91,7%.

Todos os resultados, amostras, imagens, arquivos vetoriais, códigos e materiais para validação utilizados neste trabalho estão disponíveis para visualização e possível replicação através deste link: <https://github.com/sacridini/Artigo-JGEOTEC-2020>.

CONCLUSÕES

Através dos testes realizados neste trabalho podemos verificar que a metodologia proposta apresentou resultado superior ao ser comparada a utilização da base de dados do Mapbiomas quando o objetivo é gerar camadas de florestas que não sofreram nenhuma ou pouca variação ao longo de tempo. É importante lembrar que o Mapbiomas continua sendo uma excelente referência para estudos envolvendo séries temporais, e que a comparação feita visa obter resultados melhores para uma classe que não é oficialmente mapeada pelo projeto. Os resultados obtidos poderão ser utilizados como base para outros mapeamentos de áreas florestadas estáveis em ambientes tropicais. O trabalho buscou também contribuir para um melhor entendimento das variáveis que melhor conseguem separar classes tão novas como a proposta. Além disso, resultados surpreendentes como o desempenho da banda do vermelho mostram que a utilização de apenas uma única banda pode ser o suficiente para a identificação de áreas similares e servem como um estímulo a novos estudos que visem a detecção de áreas invariantes de forma rápida e pouco custosa.

Máscaras de áreas invariantes como a que geramos poderão servir também não só como parâmetros para estudos aplicados à conservação, mas como máscaras para a limpeza de dados indesejados em resultados obtidos por algoritmos de detecção de mudanças como o Landtrendr e outros.

REFERÊNCIAS BIBLIOGRÁFICAS

- BROOKS, E. B., WYNNE, R. H., THOMAS, V. A., Blinn, C. E., and COULSTON, J. W. On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3316–3332. 2014
- BISCHL, B., LANG, M., KOTTHOFF, L., SHIFFNER, J., STUDERUS, E., CASALICCHIO, G., JONES, Z. mlr: Machine Learning in R. *Journal of Machine Learning Research*, 17(170), 1-5. 2016
- COHEN, W. B., HEALEY, P. S., YANG, Z., ZHU, Z., GORELICK, N. Diversity of algorithm and spectral band inputs improves landsat monitoring of forest disturbance. *Remote Sensing*, Volume 12. 2020
- HUANG, C., THOMAS, N., GOWARD, S. N., MASEK, J. G., Zhu, Z., TOWNSHEND, J. R. G., and VOGELMANN, J. E. Automated masking of cloud and cloud shadow for forest change analysis using landsat images. *International Journal of Remote Sensing*, 31(20):5449–5464. 2010
- JIN, S., YANG, L., DANIELSON, P., HOMER, C., FRY, J., XIAN, G. A comprehensive change detection method for updating the national land cover database to circa 2011. *Remote Sensing of Environment*, 132:159 – 175. 2013
- KENNEDY, R. E., YANG, Z., COHEN, W. B. Detecting trends in forest disturbance and recovery using yearly landsat time series: 1. landtrendr — temporal segmentation algorithms. *Remote Sensing of Environment*, 114(12):2897 – 2910. 2010
- KENNEDY, R. E., YANG, Z., COHEN, W. B., PFAFF, E., BRAATEN, J., NELSON, P. Spatial and temporal patterns of forest disturbance and regrowth within the area of the north west forest plan. *Remote Sensing of Environment*, 122:117 – 133. *Landsat Legacy Special Issue*. 2012
- KENNEDY, R. E., YANG, Z., GORELICK, N., BRAANTEN, J., CAVALCANTE, L., COHEN, WARREN B., HEALEY, S. Implementation of the LandTrendr algorithm on Google Earth Engine. *Remote Sensing*, Volume 5, Issue 5, p.1-10. 2018
- MEYER, M. C. Semi-parametric additive constrained regression. *Journal of Nonparametric Statistics*, 25(3):715–730. 2013
- MOISEN, G. G., MEYER, M. C., SCHROEDER, T. A., LIAO, X., SCHLEEWEIS, K. G., FREEMAN, E. A., Toney, C. Shape selection in landsat time series: a tool for monitoring forest dynamics. *Global Change Biology*, 22(10):3518–3528. 2016
- SOUZA, C. M., SHIMBO, J. Z., ROSA, M. R., PARENTE, L. L., ROSA, E. R. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat Data Archive and Google Earth Engine. *Remote Sensing*. 12(17). 2020
- THOMAS, N. E., HUANG, C., GOWARD, S. N., POWELL, S., RISHMAWI, K., SCHELEEWEIS, K., HINDS, A. Validation of North American forest disturbance dynamics derived from landsat time series stacks. *Remote Sensing of Environment*, 115(1):19 – 32. 2011
- VOGELMANN, J. E., XIAN, G., HOMER, C., and TOLK, B. Monitoring gradual ecosystem change using landsat time series analyses: Case studies in selected forest and range-land ecosystems. *Remote Sensing of Environment*, 122:92 – 105. *Landsat Legacy Special Issue*. 2012
- ZHU, Z. e WOODCOCK, C. E. Continuous change detection and classification of land cover using all available landsat data. *Remote Sensing of Environment*, 144:152– 171. 2014