Revisor 161958-1:

Pág 6 - Feita revisão do texto com má conectividade.

Pág. 7 – Feita revisão do parágrafo repetitivo.

Pág. 8 – Inclusão de um fluxograma para melhor entendimento dos processos de validação cruzada

Pág. 9 – Adicionadas as porcentagens de área mapeada

Pág. 10 – Com o TimeSync é possível analisar visualmente toda a série temporal de forma visual e também espectral através de um gráfico. Parte do processo (e mais demorado) de validação da ferramenta está na criação de uma série temporal anual com todos os pós-processamentos necessários para posterior análise. Com isso é possível entender se houve, por exemplo, um caso de falso positivo por conta da presença de sombras, nuvens, ou qualquer tipo de anomalia nas imagens utilizadas na série. O processo é bastante similar ao método mais tradicional de análise onde o analista classifica visualmente a amostra entre erro ou acerto. As amostras utilizadas para a validação podem ser acessadas no repositório online criado na plataforma GitHub (**https://tinyurl.com/y2kpgcvq**). O repositório também contém todos os códigos utilizados para a realização dos processos descritos no artigo, assim como todos os dados de entrada, alguns temporários importantes, resultados finais originais, mapas, amostras de treinamento e amostras utilizadas para a validação.

Revisor 162554-1:

Pág. 3 – Foram adicionadas as referências faltantes e houve correção do parágrafo para um melhor entendimento do processo descrito.

Pág. 4 – Um novo mapa para a área de estudo foi gerado com uma melhor contextualização geográfica. Houve também um aprimoramento no texto marcado como confuso pelo revisor. Além disso, realizou-se a correção da última parte do último parágrafo da introdução.

Pág. 5 – Modificações realizadas. PS: A mediana foi utilizada por ser uma medida que sofre menos influência de valores extremos. Neste caso, valores extremos são justamente valores relacionados a presença de nuvens/sombras na série.

Pág. 8 – Foi adicionado um texto explicando melhor as classes e o resultado final do mapeamento.

Pág. 8 – Visualmente a imagem não agrega muito para um melhor entendimento. No entanto, a imagem original pode ser encontrada em um repositório criado na plataforma

GitHub (**https://tinyurl.com/y2kpgcvq**). O link para o repositório foi adicionado ao corpo do texto. O repositório também contém todos os códigos utilizados para a realização dos processos descritos no artigo, assim como todos os dados de entrada, alguns temporários importantes, resultados finais originais, mapas, amostras de treinamento e amostras utilizadas para a validação.

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

Resumo

O surgimento de algoritmos de detecção de mudanças na vegetação na última década é impressionante. Mas os resultados gerados ainda possuem ruído que precisa ser tratado com a utilização de resultados de outros mapeamentos de cobertura vegetal. Além disso, a necessidade de gerar classes de uso do solo invariantes é importante para o melhor entendimento de processos que ocorrem em áreas florestais. Pensando nisso, este trabalho busca criar uma nova forma de mapear essas áreas invariáveis que possam ser utilizadas para mascarar ruídos e também como subsídio para outros estudos de conservação e restauração. A metodologia proposta usa a plataforma Google Earth Engine e um algoritmo de aprendizado de máquina: o *Random Forest*, para classificar áreas de floresta invariáveis usando todo o acervo de imagens da série temporal Landsat, de uma só vez. Os resultados mostraram que a nova abordagem teve melhor desempenho do que o uso de técnicas mais tradicionais como a agregação de mapeamentos de uso e cobertura anuais, com uma acurácia global de 91,7%. O trabalho busca ainda contribuir com a comunidade de sensoriamento remoto ao apresentar, após exaustivos testes, as melhores opções de variáveis a serem utilizadas neste tipo de classificação.

Palavras-chave: Séries Temporais, Detecção de Mudanças, Florestas, Google Earth Engine, Random Forest

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* algorithm to classify invariant forest areas using all the image's collection in the time series at once. The results showed that the new approach performed better than the use of more traditional techniques such as the aggregation of annual land-use and land-cover mappings, with an overall accuracy of 91.7%. 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: Time Series, Change Detection, Forests, Google Earth Engine, Random Forest

Resumen

La aparición de algoritmos de detección de cambios en la vegetación en la última década es impresionante. Pero los resultados todavía tienen muchos ruidos que deben ser eliminados. Además, el proceso de limpieza de datos se basa en otros mapas de cobertura de la tierra. Además de eso, es importante conocer la necesidad de generar clases de uso de la tierra invariables, particularmente en las áreas forestales. Pensando en eso, este artículo busca crear una nueva forma de mapear estas áreas invariantes que se pueden utilizar para enmascarar el ruido y como un aporte para otros estudios de conservación y restauración. La metodología propuesta aquí utiliza la plataforma Google Earth Engine y un algoritmo de aprendizaje de máquina: o *Random Forest* para clasificar áreas invariantes de bosque, utilizando a la vez todas las imágenes de la serie temporal Landsat. Los resultados encontraron que el nuevo enfoque tuvo mejor desempeño que el uso de técnicas tradicionales, con una precisión global del 91,7%. Este trabajo busca además contribuir con la comunidad de la teledetección, mostrando mediante de exhaustivas pruebas, mejores opciones de variables para utilizar en este tipo de clasificación.

Palabras clave: Series de Tiempo, Detección de Cambios, Bosques, 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 do 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, assim como de sua interpretação, agregando cada vez mais conteúdo de qualidade às análises. Alguns desses algoritmos visam a detecção de mudanças em áreas de floresta como é 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, ou seja, processos de baixa magnitude, o que possibilita estudos mais elaborados, mas que ao mesmo tempo podem gerar resultados com mais ruído.

Como se pode imaginar, a limpeza dos dados na etapa de pós processamento desses algoritmos acaba sendo 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 gerados por projetos como o Mapbiomas. Os dados do Mapbiomas podem ser utilizados como uma máscara para ignorar áreas de não interesse. Pixels que foram classificados em todos os anos do mapeamento do Mapbiomas como áreas urbanas, corpos d'água, pasto, solo exposto, agricultura e, claro, florestas, são binarizados e multiplicados entre si (classe a classe) para serem utilizados como máscaras com o objetivo de limpar possíveis ruídos no resultado final de um algoritmo como o *Landtrendr* (Kennedy et al., 2007, 2010, 2018). 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. Um pixel que possa ter sido mapeado de forma equivocada em um único ano pode anular a inclusão desta mesma área como uma área a ser utilizada como máscara. Portanto, ao utilizarmos este método, somente áreas bem consolidadas acabam sendo selecionadas para limpar o resultado gerado pelo algoritmo de detecção de mudança.

Sendo assim, o presente trabalho buscou elaborar uma alternativa a esta 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. O foco dado neste trabalho foi para a classe que chamamos aqui de florestas invariantes, já que normalmente é a classe de maior interesse e possivelmente a que possui maior quantidade de ruído associado. Diferentemente do método mais tradicional, a ideia foi realizar a criação de uma máscara gerada a partir da

classificação de toda a série temporal e não da agregação de camadas classificadas individualmente ano a ano.

Metodologia

Para a realização deste trabalho utilizou-se como área de estudo o limite a APA da Bacia Hidrográfica do Rio São João (Figura 1). A área foi escolhida por já possuir extenso número de trabalhos de referência publicados e por apresentar elementos da paisagem que representam de forma geral grande parte o contexto morfoclimático e fitogeográfico do Estado do Rio de Janeiro.

- FIGURA 1 -

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. Neste caso, floresta é tudo que foi classificado como "Formação Florestal" pelo Mapbiomas. Para isso, todas as camadas de uso do solo para o bioma da Mata Atlântica gerados pela série 4.1 do projeto (de 1985 até 2018, totalizando 33 camadas) foram recortados de acordo com os limites da APA e passaram por um processo de reclassificação, resultando em camadas binárias (floresta e não-floresta). Após a reclassificação, todas as camadas foram multiplicadas entre si resultando finalmente em uma camada binária final, onde florestas possuem com valor 1 e áreas de não-floresta valor 0.

Como a proposta metodológica consiste na aplicação de um algoritmo de classificação em toda a série temporal, foi essencial a coleta de boas amostras representativas de florestas invariantes. Boas amostras são essenciais para um resultado final satisfatório. Além disso, não seria prudente utilizar somente a camada do Mapbiomas para coletar tais amostras, já que a mesma seria posteriormente comparada com o resultado final. Sendo assim, além da utilização dos próprios dados do Mapbiomas, foi gerada uma imagem sintética do valor máximo de NDVI, considerando todos os anos da série temporal, de forma a realizar uma coleta de amostras através de uma abordagem mais conservadora.

A camada de máximo valor de NDVI foi toda desenvolvida utilizando a plataforma GEE. Para a geração 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).

Após uma filtragem inicial, aplicou-se uma função em cada uma das imagens (datas) disponíveis, gerando-se uma imagem NDVI correspondente. Através de uma outra função retirou-se todos os pixels com presença de nuvens e sombras.

Com o conjunto total de imagens NDVI, extraiu-se para cada ano da série a mediana dos valores obtidos por pixel e, a partir do novo conjunto gerado, extraiu-se o valor máximo encontrado em toda a série. Desta forma, a camada final representou o valor máximo das medianas de cada ano. A escolha da mediana como um passo intermediário foi crucial para a eliminação de ruídos e valores inválidos presentes na série temporal, dado que o cálculo baseado apenas na extração dos valores máximos gera uma camada final com ruídos espalhados por toda a imagem.

Após a criação da camada de máximo valor de NDVI, buscou-se um limiar para o valor de corte do que poderíamos classificar como 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 com a mesma metodologia realizada anteriormente para as áreas de floresta.

Tanto a camada raster de florestas invariantes (Mapbiomas) como a de pasto/agricultura invariantes (Mapbiomas) foram transformadas em pontos (vetor), selecionando-se dois mil destes pontos de forma aleatória para a aplicação de um teste t de Student. A seleção de pontos de forma aleatória se deu desta forma para evitar um possível enviesamento espacial. A aplicação do teste t de Student resultou em um valor-p de 2.2e-16, apresentando, portanto, diferença significativa entre as duas amostras. Com isso é possível concluir que as amostras com os valores de NDVI de florestas tem alta chance de pertencerem a áreas de floresta. Sendo assim, uma função para extrair o valor mínimo presente nas duas mil amostras de florestas foi processada apresentando um valor de 0,83. O valor de 0,83 pode então ser utilizado como valor de limiar de corte (*threshold*) para a geração de camadas binárias. A multiplicação da camada binária de florestas invariantes do Mapbiomas com a camada binária de máximo NDVI (todos valores maiores que 0.83) serviu então como nossa camada para extração de pontos de floresta para o treinamento do modelo de classificação.

Já para a extração de pontos de florestas que apresentaram alguma mudança significativa, utilizou-se o resultado do algoritmo *Landtrendr*. A escolha deste algoritmo

se deu à sua capacidade de detecção de eventos de baixa magnitude, o que em muitos casos significa a detecção de processos de degradação mais lentos e longos além claro da típica detecção de eventos de grande mudança. Dentre os muitos resultados que o algoritmo 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 as 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 a imagem resultante com todos os pixels que não tinham sido classificados anteriormente nem como floresta pela fusão da camada do Mapbiomas com o NDVI máximo, e 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 Earth Engine* (GEE) para execução do processo de classificação utilizando o algoritmo *Random Forest*. Como imagem base para a classificação foi necessário ter uma série temporal com imagens anuais seguindo os mesmos padrões da série do Mapbiomas para possibilitar posterior comparação. Essa série de imagens foi gerada utilizando um *script* desenvolvido no GEE contendo, para cada ano, uma composição de valores considerando o valor da mediana de cada pixel para cada banda. Para cada ano foram utilizadas as bandas Blue, Green, Red, NIR, SWIR1, SWIR2 e também de índices como o NDVI, NDWI, NDMI, SAVI, Greenness, Wetness e Brightness. Áreas com nuvens e sombra foram mascaradas com valores "no data" e, portanto, não foram consideradas 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 (*cross validation*) para entender melhor quais variáveis (bandas) teriam o melhor desempenho ao classificar este tipo de classe. Utilizamos então as amostras coletadas (250 pontos/classe) 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 GEE.

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 por teste. 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 (Figura 2). Os resultados para essa quantidade exaustiva de testes são os apresentados na Tabela 1:

- TABELA 1 –

- FIGURA 2 -

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).

Observa-se que a melhor combinação de bandas para o processo de classificação não é a simples utilização de todas as bandas. A incorporação do *Tasseled Cap* não trouxe uma melhoria na precisão do modelo. No entanto, a tabela de desempenho também apresenta outros resultados interessantes, 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, NDMI e SAVI apresentaram resultados piores do que muitas outras combinações possíveis.

Resultados

Após a otimização do modelo, foi feita uma seleção 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 3). Como resultado obtivemos uma área total de florestas invariantes na APA, de 501,1km^2.

- FIGURA 3 -

As classes escolhidas para o mapeamento final foram “Floresta Invariante”, “Perda Landtrendr” e “Outros”. A classe “Floresta Invariante” representa todos os pixels que mantiveram algum grau de coerência espectral ao longo da série temporal. Mesmo com algumas variações ao longo do tempo, o classificador pode encontrar padrões e descartar anomalias da série. Este comportamento é importante justamente para descartar ruídos presentes em qualquer série temporal sem a necessidade da aplicação de outros algoritmos de pré-processamento e por já garantir um bom resultado. Já a classe “Perdas Landtrendr” foi fortemente influenciada pelo resultado do algoritmo Landtrendr que já tinha detectado os padrões de perda (*loss*) anteriormente. Ou seja, é importante garantir que o resultado do algoritmo de detecção de mudanças seja o melhor possível modificando seus parâmetros de acordo com a área de estudo. Por fim, a classe “Outros” representa todas as outras classes presentes na bacia e que foram agregadas para não gerar confusão com os objetivos do mapeamento. O resultado da classificação final foi bastante satisfatório, apresentando um padrão espacial das classes esperado para a área de estudo.

No entanto, além da classificação gerada, uma nova imagem teve de ser criada para a comparação entre o resultado obtido pelo *Random Forest* e a imagem binária utilizando somente dados do Mapbiomas. A nova imagem foi criada somando-se a camada de florestas invariantes com todos os anos do Mapbiomas com a versão binarizada do resultado final (florestas invariantes - outros).

Com essa imagem fusionada pode se notar que grande parte dos pixels (469,8km^2 ou 76,9% da área mapeada) 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 (22,7km^2 ou 3,8% da área mapeada) e onde somente o Mapbiomas classificou como floresta (52,7km^2 ou 8,7% da área mapeada) também foram detectadas. Os outros 10,8% de área restantes foram excluídos da análise por não serem áreas de interesse.

Para entender melhor a qualidade de cada resultado, foram feitos alguns testes estatísticos. 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 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 (1949). Os resultados são os apresentados na Tabela 2:

- TABELA 2 –

Os resultados foram considerados promissores. Segundo o teste de Tukey, ao mesmo tempo em que as amostras que somente o Mapbiomas mapeou como floresta invariante são significativamente diferentes das que ambos os mapeamentos mapearam como floresta invariante. Já as amostras do *Random Forest* que são diferente das do Mapbiomas, 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 apenas nas áreas mapeadas pelo Mapbiomas ou pelo *Random Forest*, mas também nas áreas em que ambos obtiveram 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 utilização ainda consideravelmente complexa, é uma das únicas ferramentas existentes que realizam esse tipo de validação.

O TimeSync funciona de forma integrada com o GEE 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.

Foram coletados para a etapa de validação no TimeSync trezentos pontos para cada uma das classes. Os pixels mapeados apenas pelo Mapbiomas tiveram uma taxa de

acerto de 55%, enquanto os pixels mapeados 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 trezentas 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/florestas_invariantes_random_forest.

Conclusões

Através dos testes realizados neste trabalho podemos verificar que a metodologia proposta apresentou melhor desempenho se comparada à 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 já que o mesmo possui objetivos distintos. Os resultados obtidos poderão ser utilizados como base para outros mapeamentos de áreas florestadas invariantes em ambientes tropicais. O trabalho buscou também contribuir para um melhor entendimento das variáveis que melhor conseguem separar classes tão novas como a proposta, já que ainda são poucos os trabalhos que buscam classificar séries temporais nessa escala. 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 gerada pelo algoritmo de aprendizado de máquina~~,~~ poderão servir, não só como parâmetros para estudos aplicados à conservação, mas também 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. (2014) 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.

BISCHL, B., LANG, M., KOTTHOFF, L., SHIFFNER, J., STUDERUS, E., CASALICCHIO, G., JONES, Z. (2016) mlr: Machine Learning in R. Journal of Machine Learning Research, 17(170), 1-5.

COHEN, W. B., YANG, Z., KENNEDY, R. (2010) Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - Tools for calibration and validation. Remote Sensing of Environment, Volume 114, Issue 12, 2911-2924.

COHEN, W. B., HEALEY, P. S., YANG, Z., ZHU, Z., GORELICK, N. (2020) Diversity of algorithm and spectral band inputs improves landsat monitoring of forest disturbance. Remote Sensing, Volume 12.

HUANG, C., THOMAS, N., GOWARD, S. N., MASEK, J. G., ZHU, Z., TOWNSHEND, J. R. G., and VOGELMANN, J. E. (2010) Automated masking of cloud and cloud shadow for forest change analysis using landsat images. International Journal of Remote Sensing, 31(20):5449–5464.

HUGHES, M. J., KAYLOR, S. D., HAYES, D. J. (2017). Patch-based forest change detection from landsat time series. Forests, 8(5).

JIN, S., YANG, L., DANIELSON, P., HOMER, C., FRY, J., XIAN, G. (2013) A comprehensive change detection method for updating the national land cover database to circa 2011.Remote Sensing of Environment, 132:159 – 175.

KENNEDY, R. E., COHEN, W. B., & SCHROEDER, T. A. (2007). Trajectory-based change detection for automated characterization of forest disturbance dynamics. *Remote Sensing of Environment*, *110*(3), 370–386. https://doi.org/10.1016/j.rse.2007.03.010

KENNEDY, R. E., YANG, Z., COHEN, W. B. (2010) 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.

KENNEDY, R. E., YANG, Z., COHEN, W. B., PFAFF, E., BRAATEN, J., NELSON, P. (2012) 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.

KENNEDY, R. E., YANG, Z., GORELICK, N., BRAANTEN, J., CAVALCANTE, L., COHEN, WARREN B., HEALEY, S. (2018) Implementation of the LandTrendr algorithm on Google Earth Engine. Remote Sensing, Volume 5, Issue 5, p.1-10.

MEYER, M. C. Semi-parametric additive constrained regression. (2013) Journal of Nonparametric Statistics, 25(3):715–730.

MOISEN, G. G., MEYER, M. C., SCHROEDER, T. A., LIAO, X., SCHLEEWEIS, K. G., FREEMAN, E. A., Toney, C. (2016) Shape selection in landsat time series: a tool for monitoring forest dynamics. Global Change Biology, 22(10):3518–3528.

SOUZA, C. M., SHIMBO, J. Z., ROSA, M. R., PARENTE, L. L., ROSA, E. R. (2020) 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).

THOMAS, N. E., HUANG, C., GOWARD, S. N., POWELL, S., RISHMAWI, K., SCHELEEWEIS, K., HINDS, A. (2011) Validation of North American forest disturbance dynamics derived from landsat time series stacks. Remote Sensing of Environment, 115(1):19 – 32.

TUKEY, J.W. (1949) Comparing Individual Means in the Analysis of Variance. Biometrics, 5 (2): 99 – 114.

VOGELMANN, J. E., XIAN, G., HOMER, C., and TOLK, B. (2012) 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.

ZHU, Z. e WOODCOCK, C. E. (2014) Continuous change detection and classification of land cover using all available landsat data. Remote Sensing of Environment, 144:152–171.

TABELA 1: ÍNDICE KAPPA PARA AS COMBINAÇÔES 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

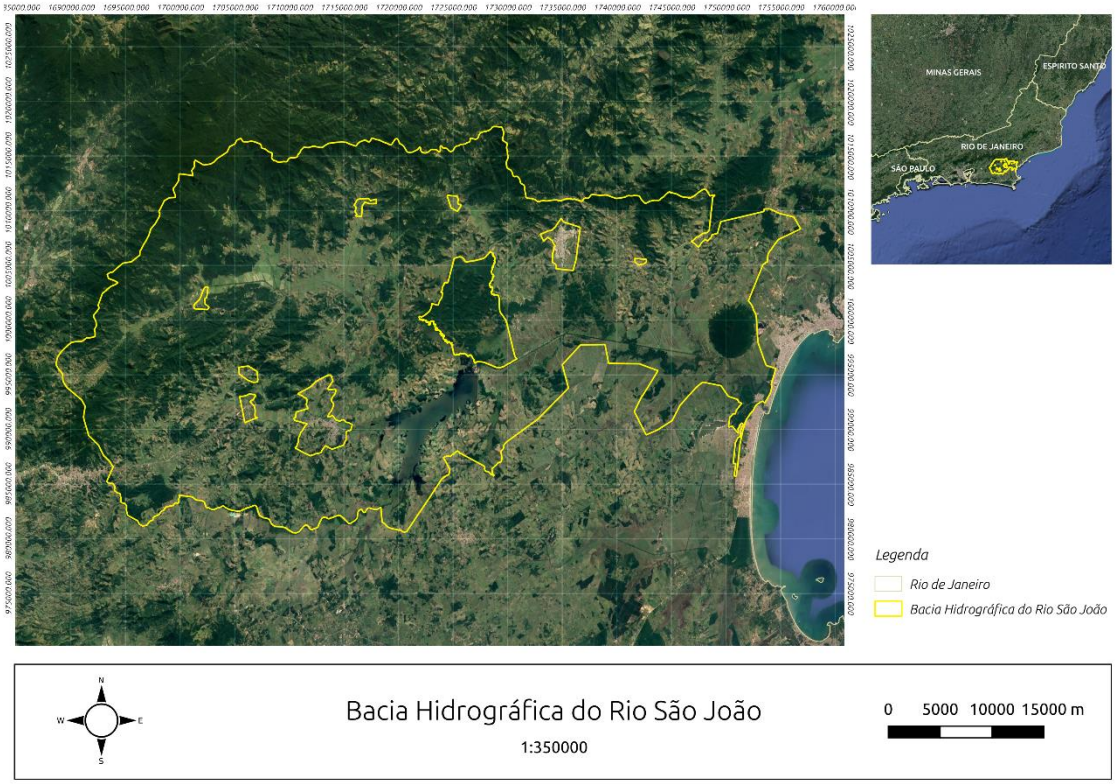

Figura 1. Área de estudo: Bacia Hidrográfica do Rio São João

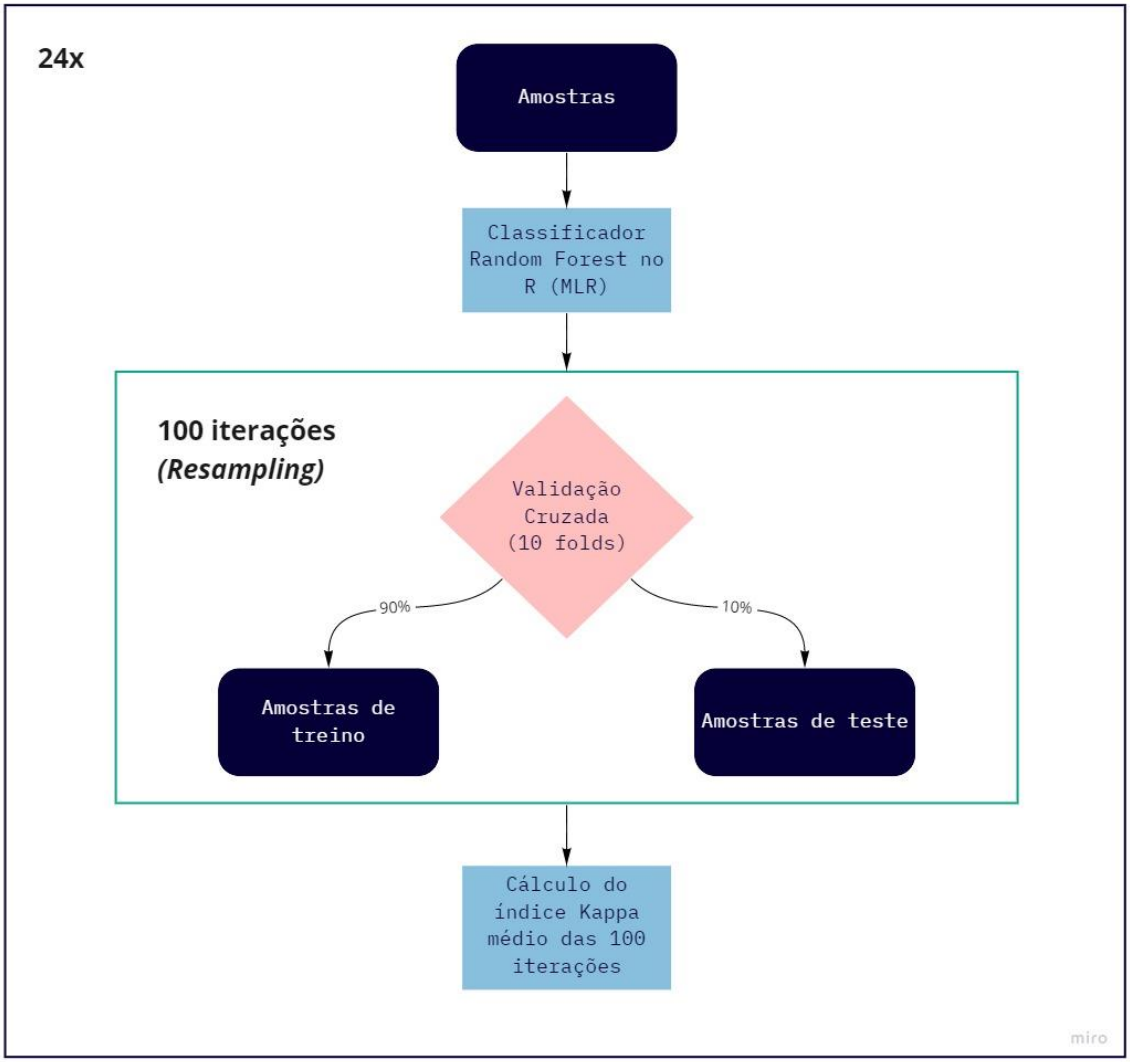

Figura 2. Fluxograma demonstrando o processo de validação cruzada. Todo o processo foi repetido para cada combinação de banda utilizada no estudo, totalizando 24 processos independentes.

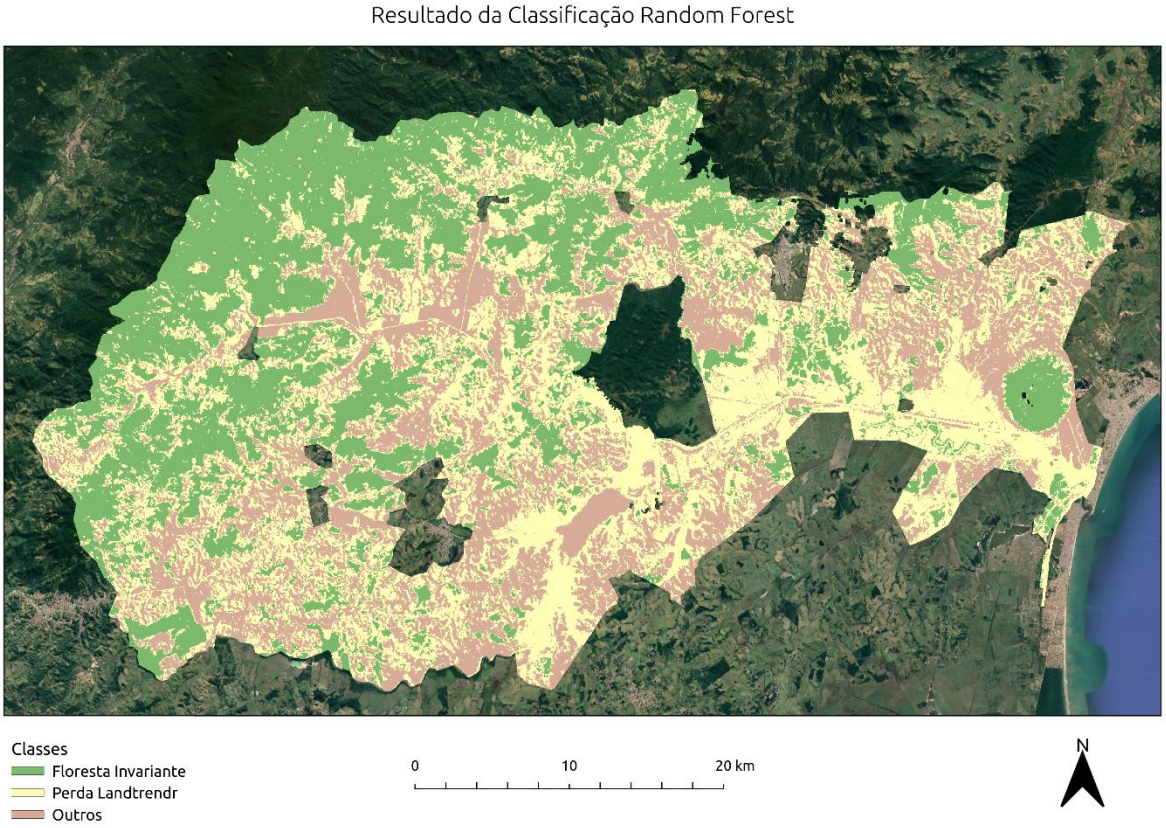

Figura 3. Classificação final gerada no Google Earth Engine (Classificação final)

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