sid.inpe.br/mtc-m21c/2020/06.30.14.25-TDI

DETECTION OF DEFORESTATION USING REMOTE SENSING TIME SERIES ANALYSIS

Alber Hamersson Sanchez Ipia

Doctorate Thesis of the Graduate Course in Earth System Science, guided by Drs. Gilberto Camara Neto and Pedro Ribeiro de Andrade Neto, approved in June 26, 2020.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/42PGNM8>

INPE
São José dos Campos
2020

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GBDIR)
Serviço de Informação e Documentação (SESID)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

BOARD OF PUBLISHING AND PRESERVATION OF INPE INTELLECTUAL PRODUCTION - CEPPII (PORTARIA Nº 176/2018/SEI-INPE):
Chairperson:
Dra. Marley Cavalcante de Lima Moscati - Centro de Previsão de Tempo e Estudos Climáticos (CGCPT)
Members:
Dra. Carina Barros Mello - Coordenação de Laboratórios Associados (COCTE)
Dr. Alisson Dal Lago - Coordenação-Geral de Ciências Espaciais e Atmosféricas (CGCEA)
Dr. Evandro Albiach Branco - Centro de Ciência do Sistema Terrestre (COCST)
Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia e Tecnologia Espacial (CGETE)
Dr. Hermann Johann Heinrich Kux - Coordenação-Geral de Observação da Terra (CGOBT)
Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação - (CPG)
Silvia Castro Marcelino - Serviço de Informação e Documentação (SESID)

DIGITAL LIBRARY:
Dr. Gerald Jean Francis Banon
Clayton Martins Pereira - Serviço de Informação e Documentação (SESID)
DOCUMENT REVIEW:
Simone Angélica Del Ducca Barbedo - Serviço de Informação e Documentação (SESID)
André Luis Dias Fernandes - Serviço de Informação e Documentação (SESID)
ELECTRONIC EDITING:
Ivone Martins - Serviço de Informação e Documentação (SESID)
Cauê Silva Fróes - Serviço de Informação e Documentação (SESID)

sid.inpe.br/mtc-m21c/2020/06.30.14.25-TDI

DETECTION OF DEFORESTATION USING REMOTE SENSING TIME SERIES ANALYSIS

Alber Hamersson Sanchez Ipia

Doctorate Thesis of the Graduate Course in Earth System Science, guided by Drs. Gilberto Camara Neto and Pedro Ribeiro de Andrade Neto, approved in June 26, 2020.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/42PGNM8>

INPE
São José dos Campos
2020

Cataloging in Publication Data

Sanchez Ipia, Alber Hamersson.

Sa55d Detection of deforestation using remote sensing time series analysis / Alber Hamersson Sanchez Ipia. – São José dos Campos : INPE, 2020.

xviii + 85 p. ; (sid.inpe.br/mtc-m21c/2020/06.30.14.25-TDI)

Thesis (Doctorate in Earth System Science) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2020.

Guiding : Drs. Gilberto Camara Neto and Pedro Ribeiro de Andrade Neto.

1. Amazon forest. 2. Deforestation. 3. Machine learning. 4. Remote Sensing. I.Title.

CDU 504.122(811)

Esta obra foi licenciada sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

Aluno (a): ***Alber Hamersson Sanchez Ipia***

Título: **"DETECTION OF DEFORESTATION USING REMOTE SENSING TIME SERIES ANALYSIS."**

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Ciência do Sistema Terrestre

Dra. Maria Isabel Sobral Escada

Mescada

Presidente / INPE / SJCampos - SP

( ) Participação por Video - Conferência

(X) Aprovado ( ) Reprovado

Dr. Gilberto Camara Neto

Mescada

Orientador(a) / INPE / SJCampos - SP

(X) Participação por Video - Conferência

(X) Aprovado ( ) Reprovado

Dr. Pedro Ribeiro de Andrade Neto

Mescada

Orientador(a) / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ( ) Reprovado

Dr. Tiago Garcia de Senna Carneiro

Mescada

Convidado(a) / ICEB / Ouro Preto - MG

(X) Participação por Video - Conferência

(X) Aprovado ( ) Reprovado

Dr. Alexandre Camargo Coutinho

Mescada

Convidado(a) / EMBRAPA / Campinas - SP

(X) Participação por Video - Conferência

(X) Aprovado ( ) Reprovado

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 26 de junho de 2020

"Science! Curse thee, thou vain toy; and cursed be all the things that cast man's eyes aloft to that heaven, whose live vividness but scorches him, as these old eyes are even now scorched with thy light".

Captain Ahab
in *"Moby Dick, Chapter 118: The Quadrant"*, 1851

ACKNOWLEDGEMENTS

I would like to thank Mom, Dad, and my brother too. I would also like to thank my co-authors and thesis directors Gilberto and Pedro and those with whom shared a place away from home: Suli, Guilherme, Leonardo, Carlos, Fabien, Mainara, Victor, Gosia, Anna, Merret, Michael, Sandra, Juliana, Mauricio, Jim, and German.

Also, I couldn't pursued a PhD without the inspiration from my friends Julio Fernando, Sergio Raul, Gregorio, Diego, Cesar, and Oscar.

This research was funded by the *"Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" - Brasil (CAPES)* - Finance Code 001, by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014-08398-6), by the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES), and the Foundation for Science, Technology and Space Applications (FUNCATE), Process 17.2.0536.1.

ABSTRACT

The Amazon rainforest plays an important role in the global carbon and water cycles, having direct influence on Earth's atmosphere and it suffers the consequences of the current climate crisis. Deforestation monitoring systems are a source of information on the forest condition for the scientific community, policy makers, and the general public. In this thesis, we identified three areas on which such systems could be improved: data processing, information extraction, and information distribution. Processing data of Earth observation satellites is subject to atmospheric noise. In particular, clouds obstruct the surveying of the Amazon rainforest. They introduce discontinuities on the the spatial and temporal patterns, which reduce the ability of analyst to extract information about features on the surface and reducing the reliability of the information obtained. Any information on Earth's surface — in our particular case, information on Land Use and Land Cover change — increases its value through sharing, validation, and reuse in broader communities. Regarding data processing, we tested several cloud detection algorithms on Sentinel-2 imagery and we found that Fmask4 provides the best performance under frequent cloud coverage. With this knowledge, we proceed to extract deforestation information using time series of the Landsat 8 and Sentinel-2 satellites, applying machine learning techniques of Deep Learning and Random Forest, respectively. We obtained the best results by using time series of Sentinel-2 images processed with Random Forest. Finally, we demonstrated the best way to provide scientists with access to massive amounts or Earth observation data and processing tools is through collaborative analysis environments offered through Internet, such as Jupyter notebooks.

Keywords: Amazon forest. Deforestation. Machine learning. Remote Sensing.

DETECÇÃO DE DESMATAMENTO USANDO SERIES DE TEMPO DE SENSORIAMENTO REMOTO NA AMAZÔNIA BRASILEIRA

RESUMO

A floresta Amazônica desempenha um papel importante nos ciclos globais de carbono e água, tendo influência direta na atmosfera terrestre e sofrendo as consequências da atual crise climática. Daí a importância dos sistemas de monitoramento de desmatamento como fonte de informação sobre a condição da floresta para comunidade científica, formadores de políticas e o público em geral. Nós identificamos três áreas nas quais esses sistemas poderiam ser aprimorados: processamento de dados, extração e distribuição de informações. O processamento de dados dos satélites de observação da Terra está sujeito ao ruído atmosférico; as nuvens, em particular, dificultam o mapeamento da floresta Amazônica. As nuvens introduzem descontinuidades nos padrões espaciais e temporais, o que reduz a capacidade dos analistas de extrair informações sobre os elementos da superfície, e também reduz a confiabilidade das informações obtidas. Qualquer informação sobre superfície da Terra, em nosso caso particular, informações sobre mudança no uso e cobertura, incrementa seu valor por meio do compartilhamento, validação e reutilização em comunidades mais amplas. Em relação ao processamento dos dados, testamos vários algoritmos de detecção de nuvens e descobrimos que o Fmask4 oferece o melhor desempenho em imagens de satélite com frequente cobertura de nuvens. Com esse conhecimento, procedemos à extração de informações sobre desmatamento usando séries temporais dos satélites Landsat 8 e Sentinel-2, aplicando as técnicas de aprendizado de máquina Deep Learning e Random Forest. Obtivemos os melhores resultados usando séries temporais de imagens Sentinel-2 processadas com Random Forest. Finalmente, demonstramos que a melhor maneira de fornecer aos cientistas acesso a grandes quantidades de dados de observação da Terra é com ferramentas de processamento e através de ambientes de análise colaborativa oferecidos pela Internet, como os notebooks Jupyter.

Palavras-chave: Floresta Amazônica. Desmatamento. Aprendizagem de máquina. Sensoriamento remoto.

LIST OF FIGURES

Page

LIST OF TABLES

Page

CONTENTS

1 INTRODUCTION

Currently, one of the few proven ways to fight the on-going climate crisis is to massively remove carbon from the atmosphere and the device we have at hand to do it are trees. However, contrary to common-sense, the trend in the tropics is towards continuing deforestation. This is disturbing, once deforestation perturbs the carbon cycle and it negatively affects water, air, biodiversity, and the sustainability of society.

Under pressure of human activities, tropical forests suffer of small scale and illegal deforestation (KALAMANDEEN et al., 2018; BRANCALION et al., 2018), fragmentation (HANSEN et al., 2020; MONTIBELLER et al., 2020), understory fire (BARLOW; PERES, 2008; BARLOW et al., 2019), agriculture expansion (GOLLNOW et al., 2018), lack of land tenure security (AZEVEDO-RAMOS et al., 2020), and infrastructure projects (BARBER et al., 2014; MUELLER et al., 2016; FEARNSIDE, 2016), among others. This coincides with the fact that such forests are continuously being considered as sources instead of sinks of CO2 (ARAGÃO et al., 2018). Despite their importance, the available global accounts of deforestation are subject to inaccuracies regarding not only their statistics but also their semantics (QIN et al., 2019; RICHARDS et al., 2017; TROPEK et al., 2014). Besides, these accounts are unfit for preventive actions because they arrive too late due to the time required to collect, organise, and process vast amounts of data.

National forest monitoring programs target specific forests aiming to prevent and discourage deforestation (CHAZDON et al., 2016; FOLEY et al., 2005; HANSEN; LOVELAND, 2012). Take, for example, two Brazilian initiatives: PRODES – Monitoring Project for Deforestation in the Legal Amazon by Satellite – and DETER – Near Real-Time Deforestation Detection System (DINIZ et al., 2015). They rely on satellite imagery, hardware, software, and human interpretation for producing medium resolution maps which played a key role during the deforestation reduction period on the first decade of the 21st century.

However, existing forest monitoring programs based on visual interpretation are close to their practical limit because of the continually increasing demands of resources considering the amount, diversity, and improvements on remote sensing data. For example, PRODES and DETER analyse the *Legal Brazilian Amazon* [1], which is 60% of the whole Amazon forest, detecting deforestation patches larger than 6.25 ha. Both

[1] In this thesis, we use the term *Amazônia* to make reference to the whole biome, while the term Amazon is used when referring to one of its parts e.g. Amazon rainforest.

projects use supervised classification methods. However, they are exclusively focused on intact forest and they cannot keep pace not only with the increasing volume of data and their finer spatiotemporal resolution, but also with emerging challenges such as forest degradation, conversion, or modification. Finally, their effectiveness is undermined by deforesters who have already learned how to avoid detection (ASNER et al., 2005; DIAS et al., 2016; SHIMABUKURO et al., 2012).

Answering the challenges posed to deforestation monitoring systems requires detailed and updated representations of Land Use and Land Cover (LUCC). Such representations would allow a deeper integration to digital representations of other components of the Earth system. With this motivation, in this thesis we consider three relevant questions: *How can satellite data can be prepared for analysis?; How best to analyse big remote sensing data sets?;* and *What are the analysis options available for LUCC scientists?*

Regarding the first question, the current satellite image distribution model is based on files covering small areas for a single date. However, for best results in analysis, Earth observation data needs to be stored using multidimensional arrays having spatial, temporal, and thematic dimensions. Thus, production of data cubes of analysis ready data (ARD) is a major trend in Earth observation. Using data cubes, processing is done in cloud services, freeing scientists from the burden of searching, gathering, and curating data. One crucial topic for producing ARD is the use of methods for detecting cloud cover and shadows. For this reason, in Chapter 3 of this document, we analyse the alternatives for cloud detection on Sentinel–2 images.

As for the second question, machine learning techniques have emerged as the primary tools for extracting information form Earth observation data cubes. However, machine learning methods are not a panacea; these methods are prone to overfitting and not good at dealing with unexpected cases. Also, many estimates of accuracy of machine learning methods for EO data analysis published in the literature rely on simple validation techniques. Usually, these validation techniques overestimate the actual performance of the classification. Therefore, it is important to be rigorous and careful when using machine learning techniques for EO data. In Chapters 2 and 4, we apply machine learning techniques for detecting deforestation in the Brazilian Amazon on Landsat 8 and Sentinel–2 images, respectively. More specifically, in Chapter 4, we undertake a careful performance assessment. This assessment shows the benefits and pitfalls of using machine learning for EO data analysis.

Finally, a remaining issue is related to the interaction of knowledge, data, and algorithms. Analysis of Earth observation data combines science and art, relying on several cycles of trial and error. To advance research, scientists rely on a hypothesis-test cycle and diaries – or *notebooks* – to keep a record of their findings. This process also relies on computer code, which scientists write their own way, following the same hypothesis-test cycle over small data sets. Nowadays, computers also help scientists to manage their digital notebooks based on concepts such as Literate Programming and Overlay Journals. These notebooks are being taken to the web in the form of collaborative analysis environments, which are on-line documents that mix code, descriptions, data, tables, and charts to summarise the results of a scientific research. This electronic approach to analysis fits well the current data distribution model based on files (KNUTH, 1984; HEY et al., 2009; PEREZ; GRANGER, 2007). In Chapter 5, we explore the properties of notebooks for analysing time series of vegetation indexes.

2 LAND COVER CLASSIFICATIONS OF CLEAR-CUT DEFORESTATION USING DEEP LEARNING[1]

2.1 Abstract

Using Deep Learning Neural Networks, we made supervised classifications of a small region of the Brazilian Amazon in order to map clear-cut deforestation. We organized Landsat 8 Surface Reflectance images into time series and we classify the images using the bands ad a Linear Mixture Model. We obtained similar accuracies using both data sets when compared to the data reported by the Brazilian Amazon Deforestation Monitoring Program (PRODES). These results suggest the possibilities of using automatic supervised techniques to extend the coverage of forest monitoring programs to those excluded areas by lack of human resources.

2.2 Introduction

Monitoring the tropical forest through remote sensing helps reducing deforestation (SEYMOUR; HARRIS, 2019). Usually, monitoring efforts focus on either accounting, alerting, or following land use after deforestation. In the Brazilian Amazon, each of these aims stands for three projects: (i) Brazilian Amazon Deforestation Monitoring Program (PRODES), whose reports accurately estimates clear-cut of pristine forest, (ii) the near real-time deforestation detection system (DETER), that produces fast alerts of change in forest areas for law enforcement authorities, and (iii) TERRACLASS, which tracks land use and cover after clear-cut deforestation (SHIMABUKURO et al., 2012).

To achieve high accuracies (e.g. TERRACLASS accuracy is above 77% (ALMEIDA et al., 2016), these monitoring projects rely on expert visual classifications, which are costly and time-consuming. For example, PRODES consolidated forest loss rates are published months after deforestation happened. In the other hand, DETER reports deforestation faster than PRODES but with lower confidence levels regarding the deforested areas. The accuracy-speed tradeoff between PRODES and DETER shapes not only their accuracy, but also the interpretation of their results. These differences make the data prone to misunderstandings by the public with daring consequences for the academia (ESCOBAR, 2019).

[1] This chapter was adapted from the proceedings of GeoInfo 2019:
Sanchez, A., Picoli, M., Andrade, P.R., Simões, R., Santos, L., Chaves, M., Begotti, R., Camara, G., 2019. Land Cover Classifications of Clear-cut Deforestation Using Deep Learning. In: Geoinfo2019. pp. 48–56.

We believe that PRODES must continue being the reference regarding deforestation in the Amazon for both historical and statistical reasons. We also believe science should explore and provide new and better answers. This brings up to the question of how to close the accuracy-speed gap by finding a cheaper and reproducible way to monitor clear-cut deforestation. An alternative is to train machine how to spot deforestation, given that they are good for boring repetitive tasks. Teaching machines is a current challenge to science and the possibility of improving forest monitoring systems with the available techniques is worth it.

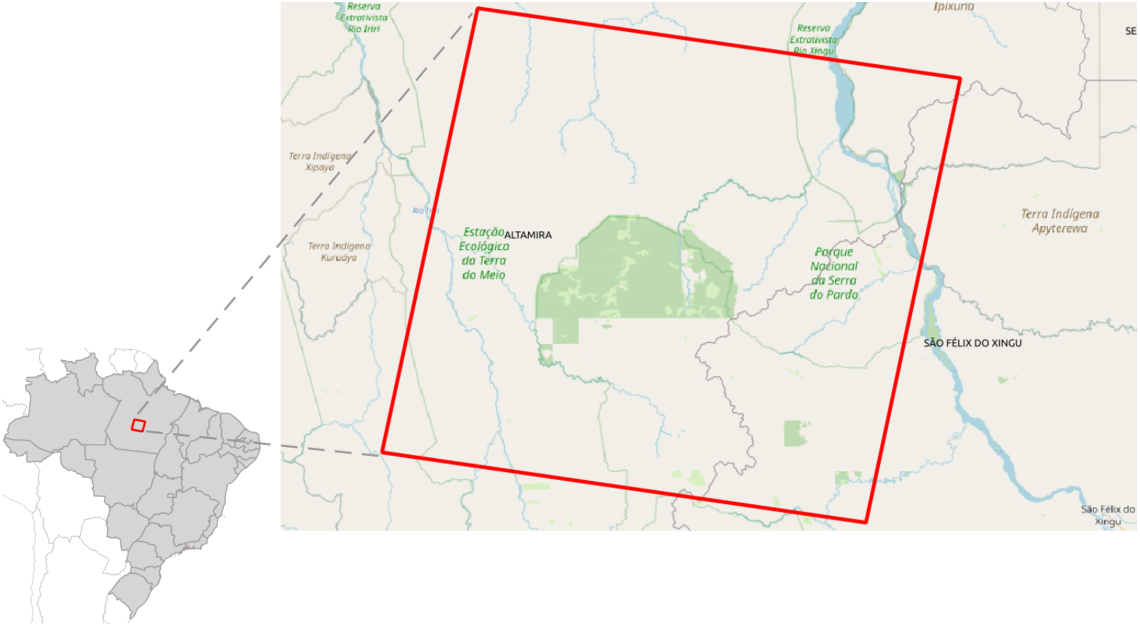
In this work, we automatically classify deforestation using Neural Networks on a study area of the Amazon rainforest. The aim of this study is to evaluate a cutting-edge classification process on deforestation detection that uses Deep Learning and satellite image time series. By comparing the raw classification maps without applying on it any post-processing algorithms, we are able to assess the bottom-line accuracy of our classification process. Our findings give us an idea on how far we are from reach the same accuracies of non-automatic visual classification systems such as PRODES. In what follows, we present the material and methods used generate the maps.

2.3 Material and methods

Our area of interest is located in the Brazilian Amazon forest, in the state of Pará, between the municipalities of *Altamira* and *São Félix de Xingu*. This area is characterized by large amounts of deforestation and a long rainy season (Figure 2.1). We obtained Landsat-8 images of the Path-Row 226/064 from National Aeronautics and Space Administration (NASA) through the Geological Service of the United States of America (WULDER et al., 2019), including atmospheric correction and cloud identification, as shown in Figure 2.2.

To train the classification algorithm, we collected sample points of forest and deforestation from the PRODES project. PRODES provides public access to deforestation data including where and when deforestation was detected. These samples were carefully selected to be representative of each class along each PRODES year (Figure 2.3 and Table 2.1).

Figure 2.1 - Area of interest. Path Row 226 064 in Landsat World Reference 2.



To prepare the data for classification, we stacked Landsat-8 images into one-year time series. We organized our data into PRODES years, which range from August to July, in order to match the seasonality of the dry and wet seasons. Each yearly dataset was stored in TIFF files, one for each variable.

Figure 2.2 - Number of clouded pixels by PRODES year from 2013 (leftmost image) to 2016 (rightmost image).

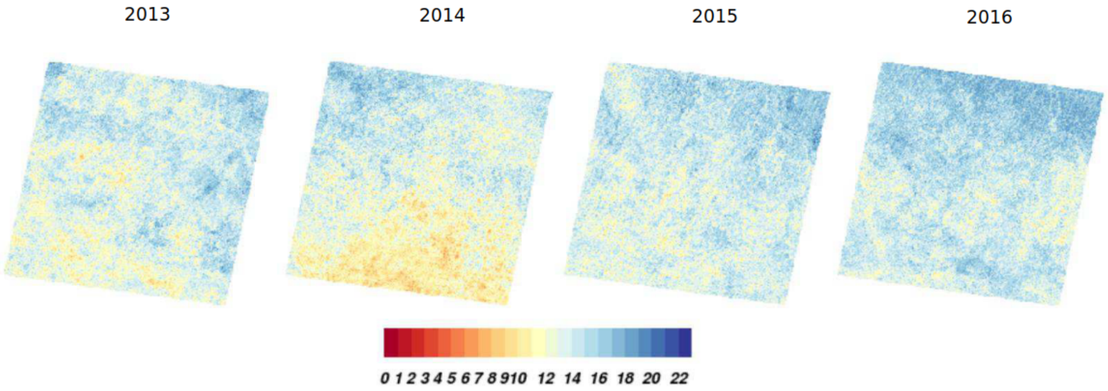
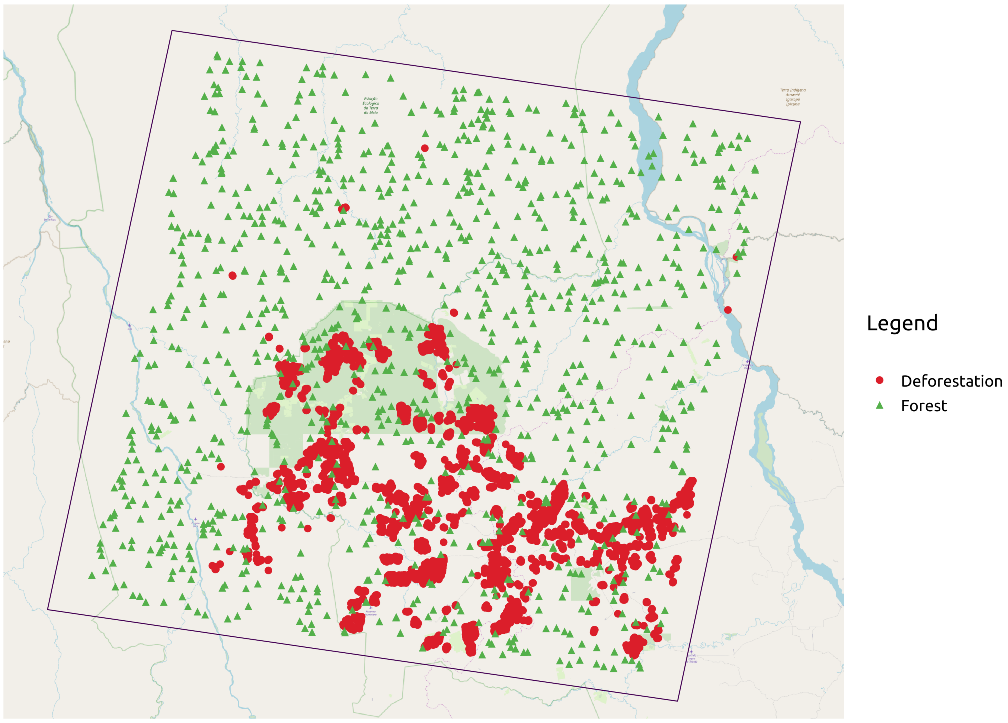


Table 2.1 - Number of samples used for training the classification algorithm.

Label	Year	Samples
Deforestation	2014	1185
	2015	1138
	2016	1122
	2017	1077
Forest	2013-2017	1100

Figure 2.3 - Sample distribution across the area of interest.



For the sake of comparison, we arranged Landsat data in three groups. The first includes Landsat bands and a vegetation index. The second includes the End Members of the global calibrated Spectral Mixture Model as described in (SOUSA; SMALL, 2017). The last one is the combination of the other two (Table 2.2).

We ran a supervised classification using Deep Learning technique. Deep Learning is concerned to statistically estimate complicated functions out of generalizable patterns in training data. This technique corresponds to supervised learning because, given a set of samples, the computer learns how to identify new (unknown) instances as forest or deforestation. As we increase the number of samples, the computer improves its classificatory capabilities (GOODFELLOW et al., 2016).

Table 2.2 - Data included in each classification.

Classification Id	Description	Variables in the classification
Bands	Landsat Bands and vegetation index.	nir, red, swir2, ndvi
MM	Spectral mixture model.	Vegetation, substrate, dark
Bands_MM	Landsat bands and mixture model.	nir, red, swir2, vegetation, ndvi, substrate, dark

We trained a feedforward Deep Learning Neural Network using the yearly time series in our samples. The training process is about finding the right parameters (weight and bias) and hyperparameters of the Neural Network. The network hyperparameters are concerned with finding the best parameters while the parameters are directly concerned in classifying the data (BENGIO, 2012). In order to maximize our chances of finding the best hyperparameters, we explored the solution space (the combinatory of all the possible hyperparameter values) by a successive process of randomization and pruning, as shown in Table 2.3.

Table 2.3 - Hyperparameter used while training our Deep Learning neural networks. All the trainings used the same optimizer (Adam), number of Layers (5), validation split (20%), and a learning rate of 0.001.

Experiment Id	Activation	Batch size	Dropout rates	Epochs	Units
Bands	selu	64	0.4	200	700
MM	selu	64	0.4	300	600
Bands_MM	sigmoid	64	0.5	300	1000

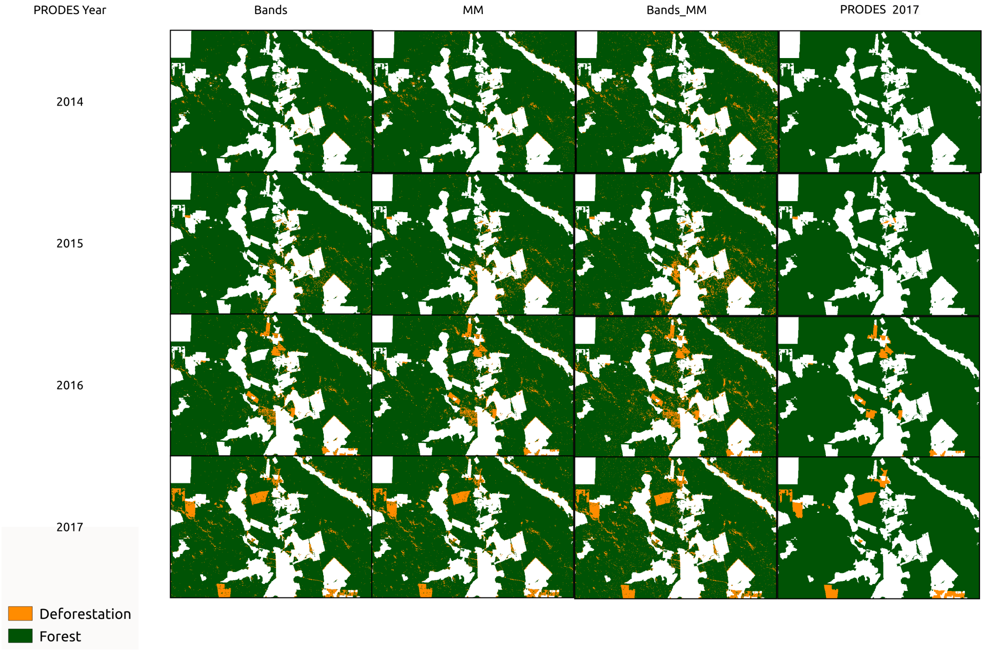
We validate our results by asking remote sensing experts to classify a set of random points, which were compared to our resulting maps. Regarding software and hardware, we used QGIS and *R* to prepare the samples, and a combination of *R*, Keras, and TensorFlow to train our neural network and to classify the images. To achieve parallelism during computations, we relied on GNU Parallel along the tools provided by operating systems based on the Linux kernel (ABADI et al., 2016; CHOLLET et al., 2017; TEAM, 2013; SIMOES et al., 2019; TANGE et al., 2011). The machine has 32 64-bit INTEL processors with 128 GB RAM running Ubuntu 14.04 with Linux kernel 4.4.

2.4 Results

Once we were done training our Network, we classify the time series derived from Landsat-8 images. We did not apply any postprocessing because we are interested in finding how far we can we reach by using only Deep Learning.

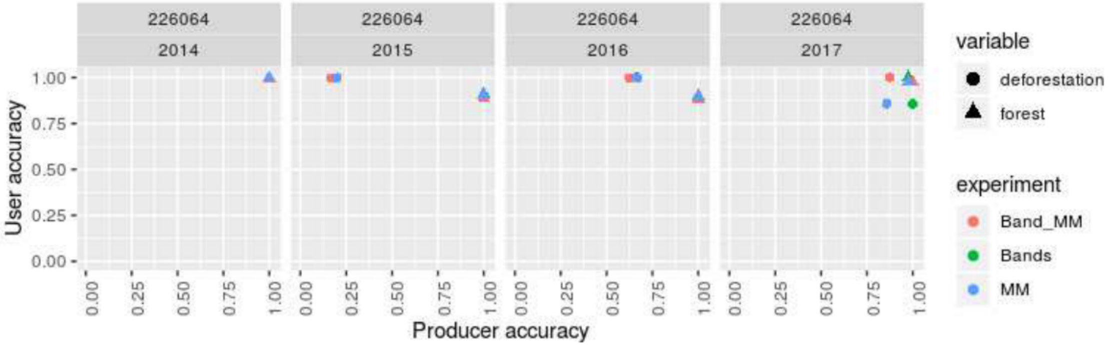
The classification results are shown in Figure 2.4. The areas painted as white are deforestation in other years, water bodies, or non-forest areas, which are ignored in the comparison. Remarkably, the classifications display small roads in the forest which are missing from the PRODES (Figure 2.4, PRODES year 2017, to the South of each map). Regarding noise, these classification presents two types: one is salt and pepper noise which is product from random errors in the classification, while the other type is elongated and clustered, resembling north-west to south-east clouds (Figure 2.4, year 2014, to the North-West and to the South-East).

Figure 2.4 - Classification results and PRODES map (right most column) from 2014 to 2017. A PRODES year runs from August to July.



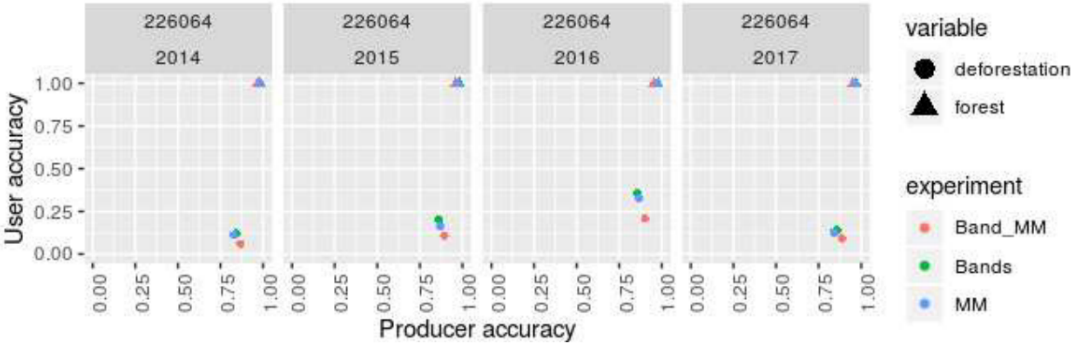
To validate our classification, we selected a set of 150 random points and then we asked experts in remote sensing to perform a visual classification. The user and producer accuracy of the classification (Figure 2.5) are above 50% with few exceptions. In general, for the forest, the producer accuracy is larger than the user and the opposite holds true for the deforestation on each PRODES year.

Figure 2.5 - Classification validation using samples classified by experts. The given years correspond to the PRODES year (August-July).



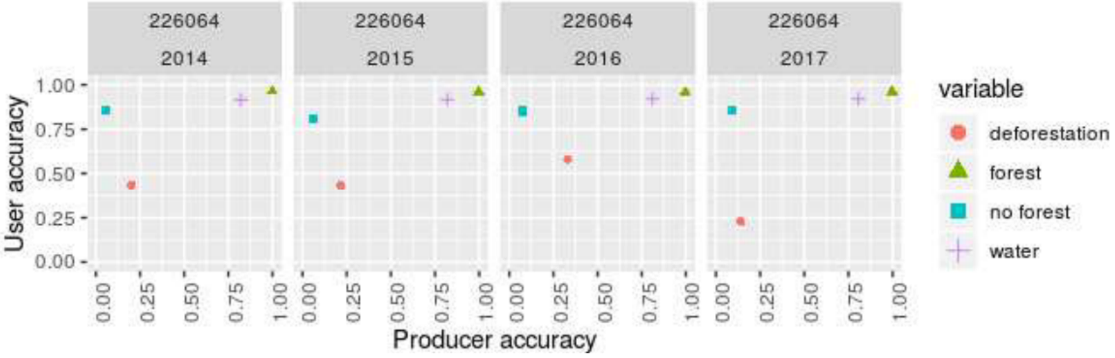
The forest validation points tend to have a producer greater than the user accuracy while the opposite holds true for the deforestation class. For the forest, this means that more often the reference data was rightly tagged. The classifier accuracy is higher for the deforestation than for the forest areas.

Figure 2.6 - Comparison of the classification to PRODES.



We also estimated how similar are our results if compared to PRODES. The similarity is reported in Figure 2.6 in the form of user and producer accuracies. While our results present large similarity regarding the forest class, for the deforestation class the user accuracy is low. As a reference, we ran the same comparison between MAPBIOMAS 4.1 (see `https://mapbiomas.org`) and PRODES and we observed high accuracy for the forest class and lower for deforestation (Figure 2.7). These results are consistent to those of (MAURANO; ESCADA, 2019).

Figure 2.7 - Comparison of MAPBIOMAS 4.1 to PRODES.



2.5 Discussion

We used annual time series of Landsat-8 data to classify a scene for the years from 2014 to 2017. Despite obtaining good classification accuracy, they are still far from those obtained by visual classification used in forest monitoring projects such as PRODES. We ran our classification using Deep Learning Neural Networks with three sets of data: Landsat bands plus NDVI, Linear Mixture Model, and their combination. However, we did not observe much difference among them in the accuracy. This is favorable for using the Linear Mixture Model giving its smaller data size and its corresponding reduction in processing time.

However, this study was constrained to a small region of the Amazon forest for short period of time. Besides, the amount of clouds in the area of interest is a limitation. Another limiting factor on the accuracy of the classifications is the relative proportions of pixels, which can induce artifacts (e.g. ratio of forest to deforestation pixels is approximately 60 to 1).

2.6 Conclusion

Monitoring the Amazon forest is hard, mainly due to its extent and almost constant cloud cover. We acknowledge this fact and at the same time reinforce the scientific need of proposing, adapting, and testing new approaches to improve classifications and/or to reduce financial costs to produce such classifications. In this work, we used Deep Learning Neural Networks over time series to identify deforestation in Landsat images. We believe that our method can support the monitoring systems because the use of time series reduces the gap between the time of deforestation and its detection.

In the results, we also found that some areas classified by us as deforestation were later found as deforestation in PRODES. We would like to quantify to which extent this corresponds to the identification of forest degradation. This is possible because PRODES only reports clear cuts. Our classifications could identify early signs of deforestation, which could improve monitoring systems as DETER.

Although the accuracy of our automatic classifications are inferior to those of visual monitoring systems such as PRODES, the approach has great potential to be improved with post-processing procedures such as spatial and temporal filters. Another possibility is to increase the temporal resolution of the images to create longer time series. A better temporal resolution might reduce the negative effects of cloudiness in our classification. To achieve this, we are planning to use products of the Harmonized Landsat Sentinel-2 project (CLAVERIE et al., 2018). Another next step in our research is to increase the area of interest to cover the whole state of Pará.

Finally, automatic classification results have the potential to help decision makers to design policies and enforce laws such as the Forest Code (Brazilian Law 12.651 of 2012). Instead of being a concurrent of visual interpretation, they can work in a complementary way. For instance, they could be used as a first step to identify deforestation using less resources if it could be possible to guarantee that false negative deforestation spots would be minimum. The errors in the automatic classifications identified visually can then be used as input to further improve the classification model.

3 COMPARISON OF CLOUD COVER DETECTION ALGORITHMS ON SENTINEL–2 IMAGES OF THE AMAZON TROPICAL FOREST[1]

3.1 Abstract

Tropical forests regulate the global water and carbon cycles and also host most of the world's biodiversity. Despite their importance, they are hard to survey due to their location, extent, and particularly, their cloud coverage. Clouds hinder the spatial and radiometric correction of satellite imagery and also diminishing the useful area on each image, making it difficult to monitor land change. For this reason, our purpose is to identify the cloud detection algorithm best suited for the Amazon rainforest on Sentinel–2 images. To achieve this, we tested four cloud detection algorithms on images spread in five areas of the Amazonia. Using more than eight thousand validation points, we compared four cloud detection methods: Fmask 4, MAJA, Sen2Cor, and s2cloudless. Our results point out that FMask 4 has the best overall accuracy on images of the Amazon region (90%), followed by Sen2Cor's (79%), MAJA (69%), and S2cloudless (52%). We note the choice of method depends on the intended use. Since MAJA reduces the number of false positives by design, users that aim to improve the producer's accuracy should consider its use.

3.2 Introduction

The world's tropical forests are essential places for environmental sustainability and the future of our planet. They combine high biodiversity and significant carbon storage with ecological services (ANTONELLI et al., 2018; OMETTO et al., 2014). Since the 1980s, the world's tropical forests have undergone substantial change. Agricultural expansion worldwide happened at the expense of tropical forest areas (GIBBS et al., 2010). In particular, the Brazilian Amazon rain forest has suffered significant deforestation. According to Brazil's National Institute for Space Research (INPE), deforestation has reached 20% of the Amazon rain forest in the country (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, 2019). Producing qualified assessments of land-use and land cover change in Amazonia is essential for evidence-based policies that can protect the forest (NEPSTAD et al., 2014; SOTERRONI et al., 2018).

Earth observation data is the primary source of assessments of deforestation in

[1] This chapter is an adapted version of the paper on MDPI's Remote Sensing journal: Sanchez, A.H., Picoli, M.C.A., Camara, G., Andrade, P.R., Chaves, M.E.D., Lechler, S., Soares, A.R., Marujo, R.F.B., Simões, R.E.O., Ferreira, K.R., Queiroz, G.R., 2020. Comparison of Cloud Cover Detection Algorithms on Sentinel–2 Images of the Amazon Tropical Forest. Remote Sens. 12, 1284.

Amazônia. Since the late 1980s, INPE produces annual estimates of clear-cut areas of forest with the PRODES project and daily indications of deforestation alerts with the DETER initiative (SHIMABUKURO et al., 2012). PRODES and DETER are the authoritative sources of information that support Brazil's actions in protecting Amazônia (ASSUNÇÃO et al., 2015; GIBBS et al., 2015a). As a companion to the monitoring of clear-cut areas, INPE and its partners also produce periodic maps of land-use change in deforested areas in Amazônia with the TerraClass project (ALMEIDA et al., 2016). Universities and research groups complement INPE's work (SOUZA JUNIOR et al., 2013; TYUKAVINA et al., 2017). Altogether, there is a substantial amount of land-use change information on the Brazilian Amazon derived from remote sensing data.

Despite the widespread data availability, the complexity of land-use transitions in Amazônia requires continuous improvements in image classification. Law enforcement actions by the Brazilian federal government managed to reduce deforestation from 2.7 Mha (million hectares) in 2004 to an average of 0.6 Mha between 2009 and 2018. Despite such reduction, deforestation in Amazonia is still at a relatively high level. To understand the complex interplay between crop production, cattle ranching, and land speculation, ever more detailed data is required.

To improve monitoring of the Amazon forest from satellites, researchers are investigating the use of big data analytics (HANSEN et al., 2013; PICOLI et al., 2018). Such methods rely on the increase of data provided by the new generation of satellites such as Sentinel–2 (DRUSCH et al., 2012). However, to use these large data sets in tropical forest areas, researchers need suitable methods of automated cloud detection in optical imagery.

Traditional alternatives for dealing with cloud cover include combining information from various dates and selecting a "best pixel" for an extended period (GRIFFITHS et al., 2013). These methods lead to the loss of temporal information required to identify crop types (BROWN et al., 2013; PICOLI et al., 2018) and pasture management (RUFIN et al., 2015; JAKIMOW et al., 2018). To capture temporal information, many researchers prefer methods that identify cloud-covered pixels and replace them with interpolated estimates (ZHU; WOODCOCK, 2012). When different satellites are combined to produce a denser time series (CLAVERIE et al., 2018), replacing cloudy pixels by interpolated values becomes feasible. For this reason, automated cloud detection algorithms are a necessary complement to big Earth observation data analytics.

Cloud detection algorithms are an active research field (ZHU; WOODCOCK, 2012; LOUIS et al., 2016; HAGOLLE et al., 2017; FRANTZ et al., 2018; QIU et al., 2019). Each algorithm has specific characteristics and ad hoc techniques; thus, comparing them on a theoretical basis is hard. In practice, performance assessment is done by selecting representative images and assessing how well each algorithm performs in each image. As an example, Baetens *et al.* (BAETENS et al., 2019) compare three cloud detection methods (MAJA, Sen2Cor, and Fmask) using 32 images from 10 different locations. As a reference for comparison, the authors use a machine learning method to identify clouds for Sentinel–2 images. Given the different land cover, the diversity of sensors, and the advances in detection methods, such comparisons serve as general guidance only.

In this paper, we approach the problem of comparing different cloud detection algorithms from a regional viewpoint. Given the importance of monitoring land change in Amazônia we consider cloud detection methods for Sentinel-2 MSI images in this region. We consider four cloud detection algorithms: Fmask 4 (QIU et al., 2019), MAJA (HAGOLLE et al., 2017), Sen2Cor 2.8 (LOUIS et al., 2016), and s2cloudless (ZUPANC, 2019). Cloud formation in Amazônia is distinct from most continental areas (ROBERTS et al., 2001). The forest produces its own rain (POSCHL et al., 2010). The rainforest generates the aerosols that make up the cloud condensation nucleus in the region (ARTAXO et al., 2009). The probabilities of cloud coverage in satellite imagery depend not only on the month of the year but also on the location inside Amazônia (ASNER, 2001). Cloud types are heterogeneous in the Amazon biome; the southern region of the Amazônia has high aerosol concentration, whereas the northern and northwestern regions have low aerosol concentration and high precipitation (CECCHINI et al., 2017; ARTAXO et al., 2009). These characteristics indicate different processes in cloud formation in subregions of the Amazon biome.

During the wet season, the precipitating clouds in the Amazon basin are either low-level stratus type clouds (up to 2–5 km altitude) or high-level convective systems (more than 6 km altitude) (ARTAXO et al., 2009). The different land cover influence the amount and type of clouds. Deep clouds are commonly found over the forest while shallow clouds are frequent over deforested areas (DURIEUX, 2003; WANG et al., 2009). Water bodies absorb visible and near-infrared radiation diminishing the reflectivity of the thin clouds above (SUN et al., 2020). The high reflectivity of artificial surfaces induces commission errors in cloud detection over urban areas (ZHU; HELMER, 2018). Such differences pose a challenge for cloud detection algorithms in Amazonia; they need to consider many types of cloud formations and associated shadows. These singular characteristics suggest that it is useful to take images over Amazônia as a study case when comparing cloud detection methods.

The rest of this article is organized as follows. We first introduce the study area and the sample regions. Then we present the cloud detection algorithms and how we configured them. Later we show how the classes resulting from each cloud detection algorithm compares to the others. Finally, we introduce our results and then we discuss some implications of this work.

3.3 Materials and methods

3.3.1 Study area

The Amazon forest covers half of Brazil (49.3%) and provides four-fifths of its groundwater (81%) with an average rainfall of approximately 2300 mm per year (DAVIDSON et al., 2012). Persistent cloud cover in Amazônia is a significant limitation for deforestation monitoring by satellite. Using the Landsat archive from 1984 to 1997, Asner (ASNER, 2001) shows how the probability of cloud cover on Landsat images depends not only on the month of the year but also on its location inside Amazônia. From June to August, the chance of finding one image with less than 30% cloud cover is 60–90% in southeastern Amazônia. In the southwestern part, cloud cover is persistent all year round. While the recent availability of medium-resolution (10–100 m) sensors with higher temporal frequency than Landsat has improved the chances of obtaining cloud-free pixels, cloud cover in rain forests such as Amazônia will always be a challenge for optical remote sensing.

3.3.2 Data selection

This study uses images from the Sentinel–2A satellite, launched in 2015. The satellite is part of the Copernicus Earth Observation program of the European Union, which is operated by the European Space Agency (ESA) and managed by the European Commission. It carries the Multispectral Instrument (MSI), which detects 13 bands of the electromagnetic spectrum spanning from the visible to the short infrared (SWIR) wavelengths at spatial resolutions of 10 m, 20 m, and 60 m, with a revisit period of 10 days (DRUSCH et al., 2012) (see Table 3.1). MSI's three bands at 60 m resolution are dedicated to atmospheric correction and cloud screening, leaving ten bands for land observation (WOLANIN et al., 2019). Sentinel-2A data enables researchers to explore the changes on Earth's surface due to its open data access policy and its temporal, spatial, and spectral resolutions.

Table 3.1 - Sentinel-2 spatial, spectral, and temporal resolution.

Band	Resolution(m)	Wavelength (nm)	Revisit period (days)
B01 Coastal aerosol	60	443	10
B02 Blue	10	490	10
B03 Green	10	560	10
B04 Red	10	665	10
B05 Vegetation red edge	20	705	10
B06 Vegetation red edge	20	740	10
B07 Vegetation red edge	20	783	10
B08 NIR	10	842	10
B8A Vegetation red edge	20	865	10
B09 Water vapour	60	945	10
B10 SWIR - Cirrus	60	1375	10
B11 SWIR	20	1610	10
B12 SWIR	20	2190	10

SOURCE: Gascon et al. (2017).

To assess cloud detection algorithms over Amazônia, we chose five areas representative of its climate heterogeneity. We identify them using the tiling system of Sentinel–2:

T19LFK: Covers part of the states of Acre and Amazonas, including an indigenous land (*Terra Indígena Apurianã*) and a protected area (*Reserva Extrativista Chico Mendes*). The region is associated with significant recent deforestation.

T20NPH: This area is in the state of Roraima, North of Brazil, and it partially covers a national forest (*Floresta Nacional de Roraima*) and an indigenous land (*Terra Indígena Yanomami*).

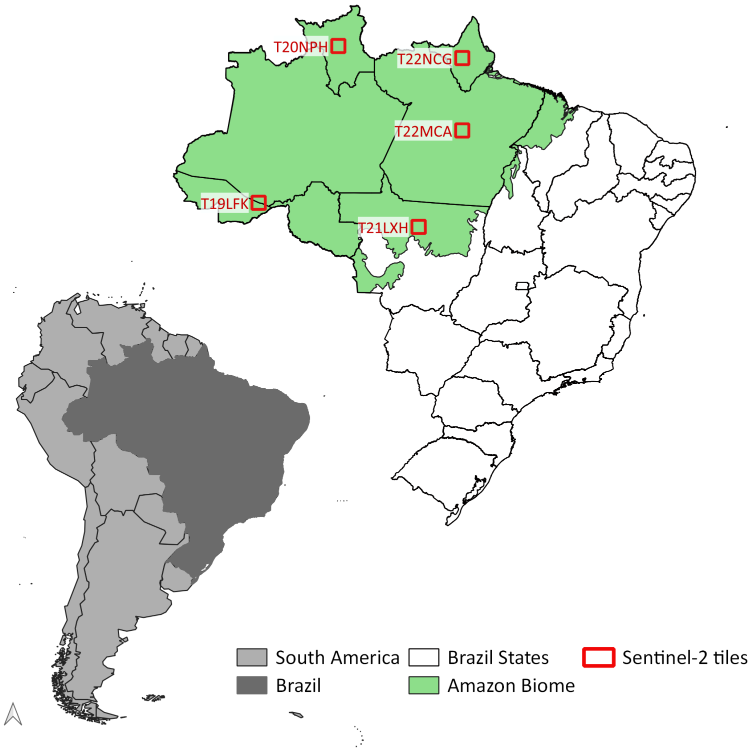
T21LXH: This area covers part of the state of Mato Grosso; it includes fragmented forest areas, soybean crops, pasture, and water reservoirs.

T22MCA: In the Para State, this area overlaps various indigenous reserves (*Arara, Araweté, Kararaô, Koatinemo, and Trincheira*) and part of a conservation unit; most of the area is covered by native forest with some deforested areas to the North.

T22NCG: This area is in the state of Amapá, including part of a National Forest (*Amapá*), a national park (*Montanhas do Tumucumaque*), and an indigenous land (*Waiãpi*).

Tiles T21LXH and T19LFK represent areas where most of the deforestation in Amazonia occurred since the 1970s. Tile T21LXH is a hotspot of Brazil's agricultural frontier with a well-defined dry season from July to September. Tile T19LFK is under the direct influence of the urban area of Rio Branco, the capital of Acre, including both deforestation and protected areas. Deforestation has increased recently in the region of tile T22MCA, threatening indigenous lands. Unlike the others, tiles T20NPH and T22NCG are in the Northern hemisphere, where the seasons and cloud patterns differ from areas to the south of the Equator. Tile T22NCG has much cloud cover all year round and low deforestation. Tile T20NPH overlaps forest and natural savanna, where emerging mining activities are menacing indigenous territories (Figure 3.1).

Figure 3.1 - Study area location. Left: Brazil in South America. Right: Location of the Amazon biome along the Sentinel–2 tiles T19LFK, T20NPH, T21LXH, T22MCA, and T22NCG.



3.3.3 Cloud detection algorithms

The paper compares four algorithms: Fmask 4 (QIU et al., 2019), MAJA (HAGOLLE et al., 2017), Sen2Cor 2.8 (LOUIS et al., 2016), and S2cloudless (ZUPANC, 2019). Fmask 4 and s2cloudless are specific for cloud detection. MAJA and Sen2Cor 2.8 are image processors; they generate cloud masks as part of image conversion from radiance at the top of the atmosphere to reflectance from ground targets. To process Landsat 8 data, USGS uses a version of the Fmask method that requires the thermal band (FOGA et al., 2017). Fmask 4 is a version of Fmask that has been adjusted to be used with sensors without thermal bands. ESA uses Sen2Cor to process Sentinel–2 images. MAJA is developed by CNES and is used by applications such as Sen2Agri (DEFOURNY et al., 2019). The Sentinel Hub uses S2cloudless for the fast generation of cloud masks (ZUPANC, 2019). These methods represent the latest generation of cloud detection algorithms for optical remote sensing images.

Fmask 4 (QIU et al., 2019) is the most recent version of Fmask (ZHU; WOODCOCK, 2012). Earlier versions of Fmask required a thermal band and worked only on Landsat images. The latest version also works on Sentinel–2 images (QIU et al., 2019). To distinguish between clouds and bright surfaces in Landsat 8 images, Fmask 4 uses the thermal band. In the case of Sentinel–2 images, it takes the view angle parallax of the NIR bands (FRANTZ et al., 2018). To reduce false positives resulting from snow and built-up areas, Fmask 4 uses spectral and contextual features. To distinguish land from water, it relies on global surface water map (PEKEL et al., 2016). Fmask 4 matches clouds with their shadows based on similarity. It iterates cloud height from a minimum to a maximum level; for each possible height, it computes the similarity between cloud and cloud shadows (ZHU et al., 2015). When processing Sentinel–2 images, its cloud and cloud shadow masks have a 20 m resolution (QIU et al., 2019).

Sen2Cor processes Sentinel–2 data to estimate Bottom-Of-Atmosphere (BOA) reflectances from Top-Of-Atmosphere (TOA) data (LOUIS et al., 2016). It takes Level-1C images and adjusts for atmospheric effects, generating Level-2A surface reflectance products (LOUIS et al., 2016; GASCON et al., 2017). It generates two types of results: (1) atmospheric correction products, such as aerosol optical thickness, surface reflectance, and water vapour maps; and (2) cloud screening and scene Classification (SCL), which assigns a class to each pixel. Sen2Cor provides two quality indicators: A cloud confidence map and a snow confidence map with values ranging from 0 to 100%. The distinction between cloudy, clear and water pixels in the SCL and the output of the cloud confidence map are used to produce the cloud confidence information (GASCON et al., 2017). The current version of Sen2Cor (2.8) increases the accuracy of classification on water, urban, and bare areas while reducing false positives for snow. Other improvements include cirrus detection, false cloud detection due to permanent bright targets, classification of water pixels inside of cloud borders, and discrimination between topographic and cloud shadow pixels (MUELLER-WILM, 2019).

Sentinel Hub's S2cloudless is a machine-learning based cloud detector (ZUPANC, 2019). Its input is Level-1C top of atmosphere data from 10 Sentinel–2 bands (bands 1-5 and 8-12) combined with pairwise band differences and band ratios. It uses the LightGBM algorithm (KE et al., 2017) trained over multiple clouded and non-clouded samples over the world. As training data, is uses cloud masks provided by MAJA as a proxy for ground truth. S2cloudless trained its classification model with $15,000$ Sentinel–2 tiles from 596 geographically unique areas in 77 different countries.

MAJA (MACCS-ATCOR Joint Algorithm) combines two methods: (a) the Multi-sensor Atmospheric Correction and Cloud Screening (MACCS); and (b) the Atmospheric & Topographic Correction (ATCOR). It builds on these methods by including time-series of images to improve detection of reflectance changes due to clouds (HAGOLLE et al., 2015). The method assumes that surface reflectances without clouds are stable in time, while clouds or cloud shadows result in quick variations (HAGOLLE et al., 2017). MAJA uses multi-temporal images that contain the most recent cloud-free observation for each pixel. At each new image, the algorithm updates this composite with the newly-available cloud-free pixels. Thus, it processes the data for a given location in chronological order (BAETENS et al., 2019). The algorithm needs to be initialized to fix cases where a given pixel has no cloud-free observations. To cover these specific cases, MAJA also uses a mono-temporal criterion based only on spectral information (HAGOLLE et al., 2017).

3.3.4 Algorithm configuration

To run Fmask 4, MAJA 3.2.2, and Sen2Cor 2.8, we used Linux Docker containers. The Fmask 4 implementation uses the MATLAB code available at GitHub (`https://github.com/GERSL/Fmask`). For Sen2Cor 2.8, we installed the version provided by ESA (`https://step.esa.int/main/third-party-plugins-2/sen2cor/`). The MAJA implementation was obtained from CNES (`https://logiciels.cnes.fr/fr/content/maja`). We downloaded S2cloudless 1.4.0 from the Sentinel Hub (`https://www.sentinel-hub.com/`). Run-time parameters are described below.

Fmask 4: Dilation parameters for cloud, cloud shadows, and snow were set to 3, 3, and 0 pixels, respectively. The cloud probability threshold was 20%, following Qiu et al. (QIU et al., 2018).

S2cloudless: Cloud probability threshold was set to 70%, using a four-pixel convolution for averaging cloud probabilities and dilation of two pixels, following the parameters set by Zupanc et al. (ZUPANC, 2019).

Sen2Cor 2.8: The tests used the same configuration as that of the Land Cover maps of ESA's Climate Change Initiative.[2]

MAJA: The evaluation used the same configuration as that of the Sen2Agri application.[3]

3.3.5 Validation sample set

To validate the resulting cloud masks, we used sample points tagged by remote sensing experts through visual interpretation, following the work of Foga *et al.* and Zhu *et al.* (FOGA et al., 2017; ZHU; HELMER, 2018). We selected a set of random locations inside each Sentinel–2 tile. Since Sentinel–2 images at 10 m resolution have over 120 million pixels, standard statistical techniques indicate that using about 400 samples per image is enough to achieve a 95% confidence level with a 5% margin of error (ISRAEL, 1992). Five experts labeled those points in 20 images of five tiles (see Table 3.2). The labels were "cloud", "cloud shadow", "clear", and "other". The "other" label is a placeholder for samples that the experts could not tag. Since the areas of cloud shadow are small compared to other labels, we tried to ensure there were at least 50 samples of cloud shadows. Two different experts classified each point; only those where both experts agreed were selected. Because of the need for agreement between experts, the final number of selected samples changes from image to image (see Table 3.2).

[2] Land Cover CCI Climate Research Data Package http://maps.elie.ucl.ac.be/CCI/viewer/download.php

[3] Sen2Agri http://www.esa-sen2agri.org.

Table 3.2 - Number of samples per image in which two experts agreed.

Tile	Date	Samples
T19LFK	2016/10/04	382
T19LFK	2017/01/02	437
T19LFK	2018/05/07	392
T19LFK	2018/11/03	452
T20NPH	2016/09/01	326
T20NPH	2016/11/10	246
T20NPH	2017/02/18	353
T20NPH	2017/07/18	311
T21LXH	2017/03/28	496
T21LXH	2018/06/11	474
T21LXH	2018/09/19	436
T21LXH	2018/10/09	457
T22MCA	2017/06/03	368
T22MCA	2017/06/23	404
T22MCA	2018/04/19	445
T22MCA	2018/06/28	447
T22NCG	2016/09/29	464
T22NCG	2016/10/19	426
T22NCG	2017/05/27	346
T22NCG	2017/07/06	433

3.3.6 Label compatibility

Since the algorithms tested (see Section 3.3.3) use different labels, we recoded their results to match the labels in the validation sample set. In particular, MAJA produces an 8-bit mask, so that many labels can be applied to a pixel, allowing combinations that are not available in the results of other algorithms. For example, MAJA's mask allows tagging a pixel as a shadow projected on top of a cloud from another cloud in a neighboring image. To make MAJA's more detailed results compatible with the output of the other methods, we prioritize clouds over cloud shadows and cloud shadows over clear pixels. Table 3.3 shows how the original codes for each method were relabelled for compatibility.

Table 3.3 - Label recoding of the detection algorithms.

Expert label	Fmask4	MAJA	s2cloudless	Sen2Cor
Clear	0 Clear land 1 Clear water 3 Snow	0-1 Clear	0 Clear	4 Vegetation 5 Non vegetated 6 Water 11 Snow
Cloud	4 Cloud	2-3 Cloud 6-7 Cloud 10-11 Cloud 14-63 Cloud 64-127 Cirrus 128-191 Cloud 192-255 Cirrus	1 Cloud	8 Cloud medium probability 9 Cloud high probability 10 Thins cirrus
Cloud shadow	2 Cloud shadow	4-5 Cloud shadow 8-9 Cloud shadow 12-13 Cloud shadow		2 Dark area pixels 3 Cloud shadows
Other				0 No data 1 Saturated or defective 7 Unclassified

3.3.7 Validation metrics

To compare the results of the algorithms, we use the F1 score (CHINCHOR, 1992) and the user's, producer's, and overall accuracies (STORY; CONGALTON, 1986). The F1 score (Equation 3.1) is the harmonic mean of the precision (Equation 3.2) and recall rates (Equation 3.3). The producer's accuracy measures how well a certain label has been classified. It is computed by dividing the correctly classified pixels in each class by the total number of pixels of the corresponding class. The user's accuracy indicates the probability that prediction represents reality. It is computed by dividing the correctly classified pixels in each label by the total number of pixels classified in that label. The overall accuracy indicates the quality of the map classification. It is calculated by dividing the total number of correctly classified pixels by the total number of reference pixels.

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.1}$$

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.2}$$

$$Recall = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{3.3}$$

3.4 Results

In our experiments, Fmask 4 has the best overall accuracy, followed by Sen2Cor, MAJA, and S2cloudless (see Table 3.4). Fmask 4 consistently outperforms the other algorithms in overall, user's and producer's accuracies for all classes. It also has the best results considering individual tiles. For cloud shadow detection, Fmask 4 has a better performance than Sen2Cor. Although MAJA has the ability to detect cloud shadows, in practice the methods extend its cloud mask to include shadows. MAJA is a conservative method that uses dilation operators to improve the user's accuracy of the clear sky. Therefore, no shadows are reported by MAJA. This is observable in either the images themselves or Figure 3.2, in which MAJA consistently detects more cloud pixels than the other methods. For accuracy assessment, we merged both types of clouds for computing the information in Table 3.4 and Table 3.5 to be able to compare Fmask 4 and Sen2Cor with the other methods.

Table 3.4 - F1 scores, user, and producer accuracies for each cloud detection algorithm.

	Fmask4			MAJA			s2cloudless			Sen2Cor		
Label	F1	user	prod	F1	user	prod	F1	user	prod	F1	user	prod
Clear	0.90	0.90	0.89	0.73	0.82	0.66	0.44	0.42	0.46	0.77	0.67	0.89
Cloud	0.94	0.91	0.96	0.77	0.64	0.97	0.66	0.59	0.75	0.89	0.90	0.88
C. Shadow	0.79	0.84	0.75			0.00			0.00	0.50	0.95	0.34
Overall		0.90			0.69			0.52			0.79	

When comparing the overall accuracy of Sen2Cor with that of MAJA, their design choices stand out. MAJA has a better user's accuracy for clear sky pixels than Sen2Cor; for producer's accuracy of this class, Sen2Cor is superior. Conversely, MAJA has a better producer's accuracy than Sen2Cor for cloud pixels; for the user's accuracy, the situation is inverted. This situation also holds in individual tiles (see Table 3.5). The designers of MAJA have chosen to maximize the probability that pixels labeled as a clear sky are correct.

The behavior of S2cloudless is erratic; sometimes it produces results visually similar to those of Fmask 4 or Sen2Cor (see Figure 3.3), while in some other occasions it misclassifies clear pixels as clouds (see Figure 3.2). For example, for tile T21LXH on 28 March 2017 and tile T22MCA on 28 June 2018, S2cloudless has a particularly poor performance. Figure 3.4) shows tile T21LXH on 28 March 2017, a case where S2cloudless performs differently from the other methods.

These results showed that the four algorithms produce their best results on images with few well-defined (crisp) clouds. Except for s2cloudless, the algorithms agree on the shape and the number of areas classified as either cloud or clear. However, they cannot adequately approximate the shape of clouds and their shadows on thin semi-transparent cirrus or tightly packed clouds (see Figure 3.5). The accurate detection of cloud shadows is challenging because dark surfaces, such as wetlands, burned areas, and terrain shadows can be easily confused with cloud shadows (QIU et al., 2019).

Table 3.5 - User and producer accuracies for each tile and cloud-detection algorithm.

		Fmask4			MAJA			s2cloudless			Sen2Cor		
Tile	Label	F1	user	prod	F1	user	prod	F1	user	prod	F1	user	prod
T19LFK	Clear	0.83	0.81	0.86	0.66	0.69	0.63	0.47	0.31	0.94	0.66	0.52	0.92
	Cloud	0.96	0.96	0.96	0.90	0.85	0.97	0.77	0.94	0.66	0.94	0.96	0.92
	C. Shadow	0.68	0.71	0.66			0.00			0.00			0.00
	Overall		0.92			0.82			0.64			0.84	
T20NPH	Clear	0.91	0.95	0.88	0.78	0.89	0.70	0.53	0.47	0.62	0.84	0.73	1.00
	Cloud	0.95	0.90	1.00	0.71	0.56	0.98	0.57	0.54	0.60	0.93	0.99	0.88
	C. Shadow	0.80	0.82	0.78			0.00			0.00	0.59	1.00	0.42
	Overall		0.91			0.67			0.50			0.84	
T21LXH	Clear	0.88	0.82	0.95	0.80	0.89	0.72	0.35	0.33	0.36	0.78	0.64	0.99
	Cloud	0.94	0.96	0.92	0.77	0.64	0.98	0.60	0.52	0.70	0.91	0.99	0.83
	C. Shadow	0.81	0.89	0.75			0.00			0.00	0.58	0.98	0.41
	Overall		0.90			0.71			0.45			0.81	
T22MCA	Clear	0.94	0.94	0.94	0.88	0.83	0.93	0.58	0.62	0.54	0.85	0.74	0.98
	Cloud	0.94	0.90	0.98	0.82	0.71	0.97	0.70	0.56	0.93	0.95	1.00	0.90
	C. Shadow	0.81	0.89	0.74			0.00			0.00	0.49	0.87	0.34
	Overall		0.92			0.77			0.58			0.83	
T22NCG	Clear	0.87	0.95	0.80	0.42	0.70	0.30	0.23	0.34	0.18	0.63	0.64	0.63
	Cloud	0.87	0.79	0.96	0.58	0.43	0.92	0.61	0.46	0.94	0.71	0.62	0.83
	C. Shadow	0.80	0.82	0.77			0.00			0.00	0.53	0.97	0.36
	Overall		0.86			0.48			0.43			0.65	

Figure 3.2 - Number of pixels classified on each Sentinel–2A image.

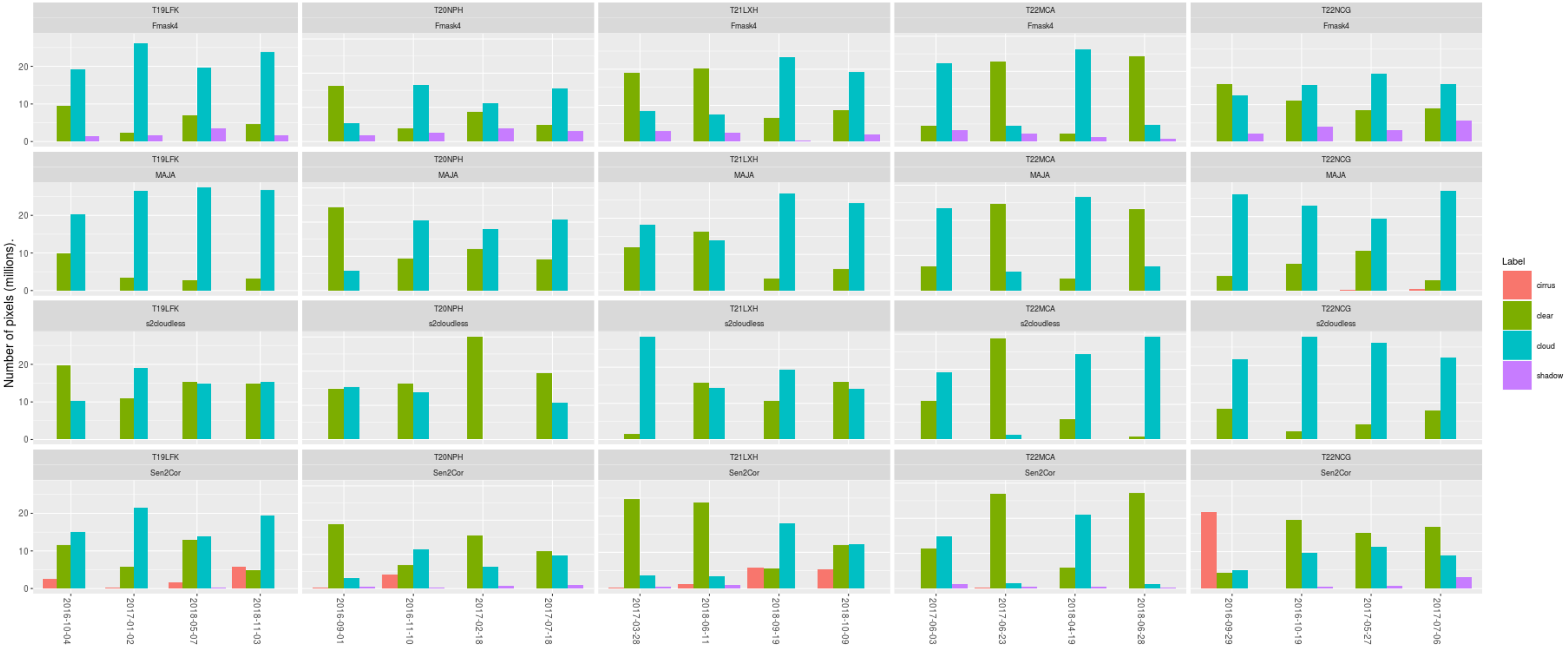
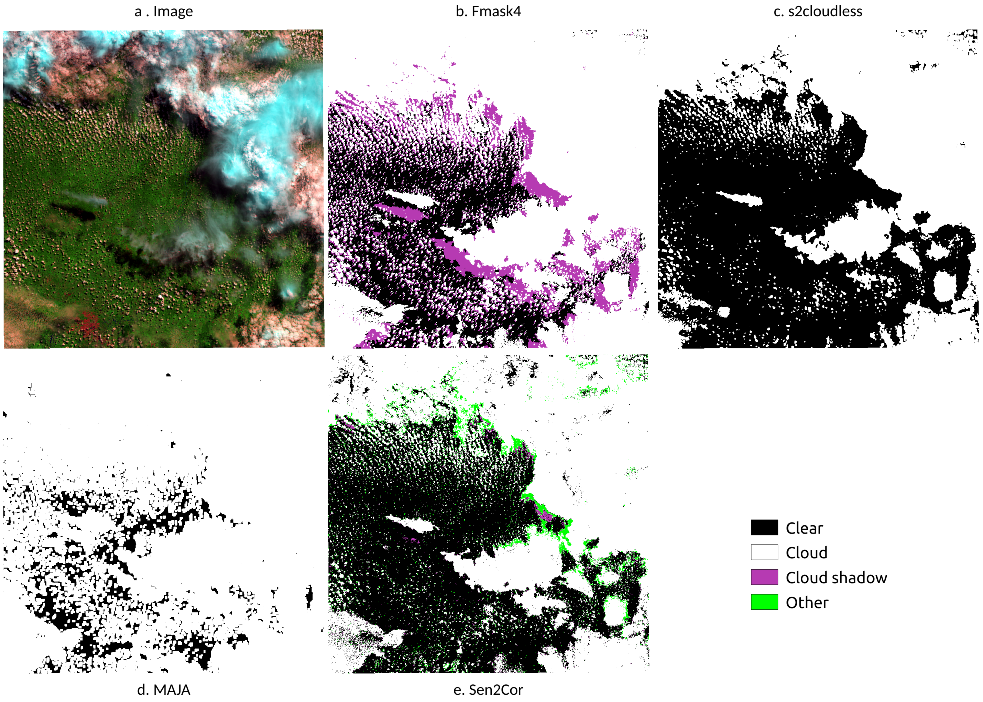
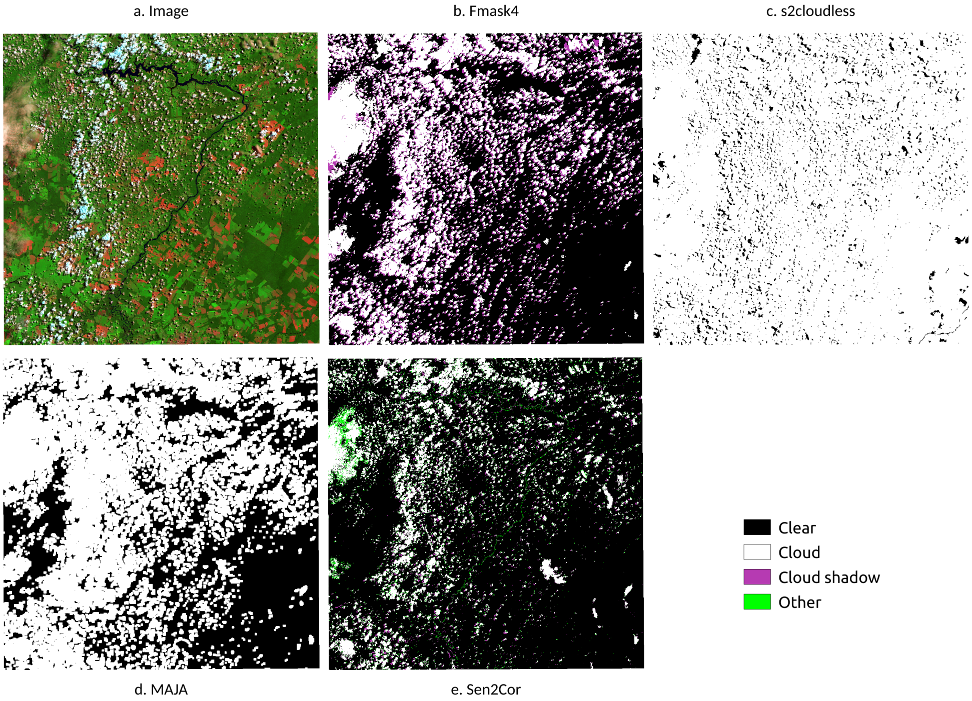


Figure 3.3 - Clouds detected on the Sentinel–2A image T19LFK of May 7, 2018. The color picture (a) uses bands 12, 8 and 3.



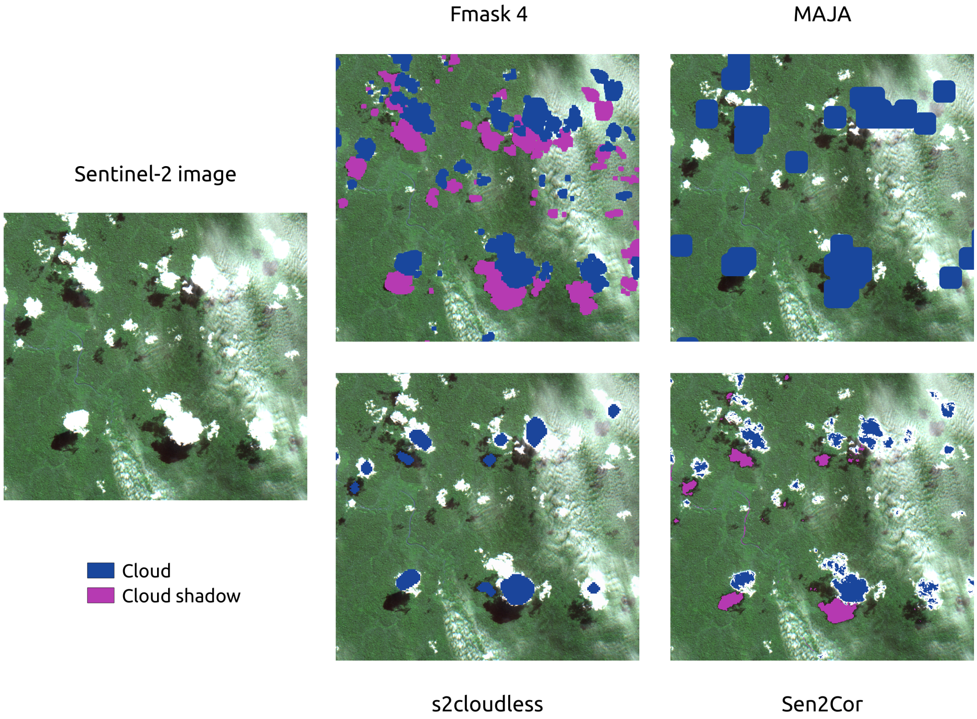
The pictures shown in Figure 3.4 confirm the results discussed above. While Fmask4 has the best performance, it is interesting to compare MAJA with Sen2Cor. MAJA uses squircles (*i.e.*, a shape between a circle and a rectangle) to fill in the cloud shape, ensuring the total coverage of each cloud in detriment of cloud shadows. Thus, MAJA sometimes incorporates clear pixels in its cloud mask, as also reported by (BAETENS et al., 2019). By contrast, Sen2Cor approximates the shape of the cloud from the inside, filling in the clouds' boundaries with saturated labels — particularly with thin cirrus clouds — which produces rough borders (see Figure 3.5). Furthermore, Sen2Cor cannot detect small clouds which are correctly identified by Fmask 4 and MAJA (*e.g.* see the small clouds at the center of Figure 3.3).

Figure 3.4 - Clouds detected on the Sentinel–2A image T21LXH of March 28, 2017. The image is composed of bands 12, 8 and 3.



The shadow masks produced by Fmask 4 are displaced regarding to the shadows visible in the images. This is a consequence of the coarse spatial resolution of the digital elevation model used by the method. As for their shapes, Fmask 4 matches well the cloud shadows respect to the clouds producing them, while in Sen2Cor the cloud shadows have smoother and different boundaries than their clouds (see Figure 3.5). Our results confirm the work of Qiu *et al.* (QIU et al., 2019) and Baetens *et al.* (BAETENS et al., 2019), who report that Fmask 4 works better in detecting clouds and cloud shadows than Sen2Cor for Sentinel–2A images.

Figure 3.5 - Detail of clouds over Amazonia in Sentinel–2A image T22MCA of June 23, 2017. The image is composed of bands 4, 3 and 2.



We could not compute the accuracies for cloud shadow detection for MAJA and S2cloudless. This is expected from s2cloudless but it comes as a surprise in the case of MAJA. An explanation is MAJA's greedy behavior regarding clouds; it tends to tag pixels as clouds in disregard of their shadows (see Figure 3.2). An alternative explanation is due to our interpretation of the MAJA's bit mask where we prioritized clouds over cloud shadows (see Section 3.3.6).

Sen2Cor tags many pixels as saturated, defective, or unclassified which we labeled as *other* (see Table 3.3). A visual inspection reveals most of the saturated pixels are the external border of clouds (see Figure 3.3). On the other hand, Sen2Cor in tile T21LXH for 08 March 2017 (see Figure 3.4) mistakenly displays cloud and cloud shadow pixels along the riverbank, almost perfectly profiling the whole river; this could be caused by suspended matter in the water. Sen2Cor problems to detect cloud cover over water were also reported by (QIU et al., 2019; SEGAL-ROZENHAIMER et al., 2020). As discussed above, the shapes of clouds in the Sen2Cor mask are rougher compared to the smoother results using the other algorithms. Despite these issues, Sen2Cor is a reliable method for cloud detection. Its producer's accuracies for clear sky and clouds are respectively 89% and 88%. If its errors in detecting cloud shadows can be tolerated by the user, its efficiency and ease of use may justify its choice for bulk processing.

S2cloudless erratically mixes land features and clouds, particularly on images with few clouds. It does not spot cloud shadows. We could not to confirm the claims made by the authors of this method (ZUPANC, 2019) about the good performance of this algorithm. One explanation is that the clouds in tropical forests such as Amazonia are not adequately included in the S2cloudless training set.

3.5 Discussion

The results of this study show that the Fmask 4 algorithm consistently performs better than the alternatives for Sentinel-2 images of the Amazon rain forest. Fmask 4 had an overall accuracy of 90%, followed by Sen2Cor (79%), MAJA (69%) and s2cloudless (52%). Our results are different from those of Baetens et al. (BAETENS et al., 2019) who concluded that MAJA and Fmask 4 perform similarly with an overall accuracy around 90%, while Sen2Cor had an overall accuracy of 84%. We now consider some hypothesis that could account for such significant differences.

As noted by Baetens et al. (BAETENS et al., 2019), for satellites without thermal bands cloud detection methods use thresholds. Different thresholds are set for the visible bands, the 1.38 µm band, and the Normalized Difference Snow Index. These approximations address important challenges for cloud detection methods: distinguishing clouds from snow, mountain tops, bright deserts, and large built-up objects. Since each cloud detection method relies on different *ad hoc* hypotheses, its usefulness varies from scene to scene. For this reason, no single study can provide definitive guidance. Studies that target specific regions, such as the current paper, provide valuable advice even though its results cannot be generalized to non-forest areas.

A comparison done by Baetens et al. (BAETENS et al., 2019) uses 10 different sites, including equatorial forests, deserts and semi-deserts, agricultural areas, mountains, and snowy areas. Their results provide a balance between different targets that could be confused with clouds. By contrast, our study deals only with forest and agriculture areas; the images tested have no deserts, mountains or snow. By focusing on the Amazon biome, our results are intended as guidance for experts interested in measuring land change in the region. Given its focus, these results cannot be generalized to non-forest regions.

A further consideration that could explain part of the differences between our work and that of Baetens et al. (BAETENS et al., 2019) is the choice of training data sets. While we use random sampling, those authors preferred to rely on active learning. An active learning model uses a few good quality samples instead of a large ensemble of random points. These good quality samples are used to train a machine learning model (random forest) whose output provides labels to a large set point for classification. In theory, this method has the advantage of being able to provide a larger number of points to test the algorithm. Machine learning models have a tendency to overfit their training data, which could cause wrong predictions (HASTIE et al., 2009). The alternative is to use random samples, which rely on standard statistical assumptions. However, random samples can miss some cloud properties. Clouds come in different shapes, sizes, and transparencies; it is often hard to distinguish overlapping clouds at different heights from images. Random samples can also misrepresent minority labels such as cloud shadows. Therefore, both random sampling and active learning have their advantages and shortcomings for evaluating cloud detection algorithms. Further testing and comparison are required to evaluate these approaches.

Another source of divergence between our result and that of Baetens et al. (BAETENS et al., 2019) is due to class relabeling. Cloud detection algorithms codify their results using different levels of detail. To enable comparisons, we had to recode them to the same set of labels. This process implies a loss of information, in particular for MAJA, which provides the most detailed data about its detection process. Thus, our recoding process could have had a negative impact on our evaluation of MAJA.

Despite the differences discussed above, there are points of convergence between our work and earlier papers such as Baetens et al. (BAETENS et al., 2019) and Qiu et al. (QIU et al., 2019) related to Fmask 4 performance. The overall user's and producer's accuracy values for Fmask 4 are broadly consistent in the three studies. Qiu et al. (QIU et al., 2019) report producer's accuracies for clouds, shadows and clear pixels to be 93%, 70%, and 97%, while our results are 96%, 75%, and 90%. Thus, we consider that Fmask 4 to be a reliable method that we recommend to be used for cloud detection in Sentinel-2 images of the Amazon rain forest.

3.6 Conclusion

In this work, we compared four cloud detection algorithms on Sentinel–2A images of the Amazon tropical forest, and we found that Fmask 4 performs the best. We tested four cloud detection algorithms — FMask 4, Sen2Cor, MAJA, and S2cloudless — on 20 images with a different amount of cloud coverage, spread over five regions of Amazonia. We validated the results of the cloud detection algorithms using the criteria of experts on remote sensing who classified approximately 400 random points on each image. To determine the best algorithm, we computed the F1 score and the overall, user, and producer accuracies. We found that FMask 4 has an overall accuracy of 90% to detect clouds, while Sen2Cor's OA is 79%, MAJA's OA is 69%, and S2cloudless's OA is 52%. Based on these results, we recommend the use of Fmask 4 for cloud detection of Sentinel-2 images of the Amazon region.

The choice of method depends on the intended use. Therefore, users should consider the benefits of each method before making their choices. Since MAJA reduces the number of false positives by design, users that aim to improve the producer's accuracy should consider its use. These characteristics could make MAJA suitable, for example, to build cloud-free monthly mosaics. Despite the poor performance of S2cloudless in our study, we consider that the use of machine learning methods for cloud detection is a promising way forward. As more good quality samples become available, its performance will improve. Finally, Sen2Cor is an efficient method to detect clouds in Sentinel-2 images. Despite not having the best performance, its ease of use may appeal to those that need fast processing of large data sets.

We expect our work to impact on the building of data cubes of analysis-ready data from satellite imagery, like those currently under construction by the Brazil Data Cube project[4]. Another application is for improving the time series analysis of Land Use and Land Cover change of deforested areas, which is particularly hard because of cloud coverage. Given the performance of FMask 4, space agencies and committees such as CEOS should consider the value of working together to develop a standardized best quality cloud detection methods that could be shared and used for remote sensing optical imagery. The *R* and *Python* scripts used to compare the performance of cloud detection algorithms are available on GitHub: https://github.com/brazil-data-cube/compare-cloud-masks.

[4] Brazil Data Cube project http://brazildatacube.org/

4 DETECTING TROPICAL DEFORESTATION USING TIME SERIES OF SENTINEL-2A IMAGES[1]

4.1 Abstract

In this paper, we use a machine learning algorithm combined with satellite image time series at 10 meter resolution for detecting small deforestation patches in the Amazon rainforest. We ran a supervised classification using random forest and Sentinel-2A images for identifying Deforestation, Forest and Other classes. For classifying, we selected two different sets of attributes, one including only bands and another including exclusively vegetation indices. We used 36 images from August 2018 to July 2019 of the tile 20LKP, located in the border between the west of Brazil and the north of Bolivia. The images were processed to surface reflectance and the clouds were masked using the algorithm Fmask. Then, we used K-Fold technique for selecting the best combinations of Sentinel-2A bands and vegetation indices. Later, we used bootstrapping for improving the good practices for accuracy assessment. Our classification using bands obtained F_1 score of 93% for Deforestation, 85% for Forest, and 78% for Other while using vegetation indices we obtained 91%, 85%, and 82% for Deforestation, Forest, and Other, respectively. These results indicate that our proposed method is scalable and accurate, which is of paramount importance for forest monitoring systems that support decision and policy makers in the context of the current global climate crisis.

4.2 Introduction

Forests play an important role in the global climate by regulating the water and carbon cycles. However, deforestation diminishes the capacity of the forest to store carbon, putting it back into the atmosphere, worsening the current climate crisis (EXBRAYAT et al., 2017). To protect the Amazon tropical forest, Brazil set up law enforcement policies and monitoring systems, which reduced deforestation during the first decade of the 21st century. However, this decreasing trend has been reverted in the last years (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, 2020). This new trend can be explained by a direct expansion of pasture areas, an indirect expansion of agricultural crops, a relaxation of laws protecting indigenous lands, and a lack of policy enforcement (PICOLI et al., 2020; RICHARDS, 2015; BEGOTTI; PERES, 2020; ANONYMOUS, 2019).Additionally, the current Brazilian government has allowed farmers to grow sugar cane in the Amazon by revoking in November of 2019 the Decree 6,961 of 2009 (FERRANTE; FEARNSIDE, 2020), and it is presenting a

[1] This chapter is being submitted to the MDPI's Remote Sensing journal.

bill to the Congress allowing mining exploration in indigenous lands (bill 191/2020).

Current deforestation monitoring systems heavily rely on Remote Sensing platforms for data acquisition (FINER et al., 2018). These systems provide the public, decision makers, and law enforcement agencies with either deforestation accounting, deforestation warnings, or both (ACHARD et al., 2010; MILODOWSKI et al., 2017). Examples of the former are PRODES and Global Forest Watch (GFW) (SHIMABUKURO et al., 2012; HANSEN et al., 2013), and of the latter are DETER (DINIZ et al., 2015), MapBiomas Alert[2], and Imazon Deforestation Alert System (SAD)[3]. Deforestation monitoring systems such as GFW and PRODES use 30 meter resolution images from the Landsat program. This program has been collecting satellite imagery since the 1960s, which fits well for deforestation accounting and long-term trend analysis (WULDER et al., 2019). However, the rise of small area deforestation challenges current monitoring systems, implying a need for constantly improving these systems in regards of their resolution, coverage, and accuracy (HANSEN et al., 2020; KALAMANDEEN et al., 2018; RICHARDS et al., 2017).

Cloud computing services are becoming cheaper, enabling access to a large amount of processing power and storage capacity. These advances allow scientists to process larger Earth Observation (EO) data at finer resolutions and at the same time covering larger extents (GIULIANI et al., 2019; MAHECHA et al., 2020). However, we still need effective methods to improve accuracy, accuracy assessment, and take advantage of all available data in an automated process.

Most Land Use and Land Cover change maps using remote sensing published in the literature use the space-first, time-later paradigm, also known as multitemporal classification (CAMARA et al., 2016a). These maps compare changes between two dates (multitemporal images); however, such two-date comparisons miss the actual change information, which is only available in the temporal component. This deficiency can be tackled by employing time series, which can better characterize the phenomena on Earth's surface by describing both trends and discrete events of change (GOMEZ et al., 2016). Moreover, advances in machine learning techniques can deal with the large volumes of data available nowadays. Machine learning algorithms have been used for Land Use and Land Cover classification (PICOLI et al., 2018; SIMOES et al., 2020) and for deforestation detection (GRINAND et al., 2013; ADARME et al., 2020). Time series combined with machine learning techniques might

[2] MapBiomas Alert is available at `http://alerta.mapbiomas.org`

[3] Imazon SAD is available at `https://infoamazonia.org/en/datasets/deforestation-imazon-sad/`

produce high accuracy classifications.

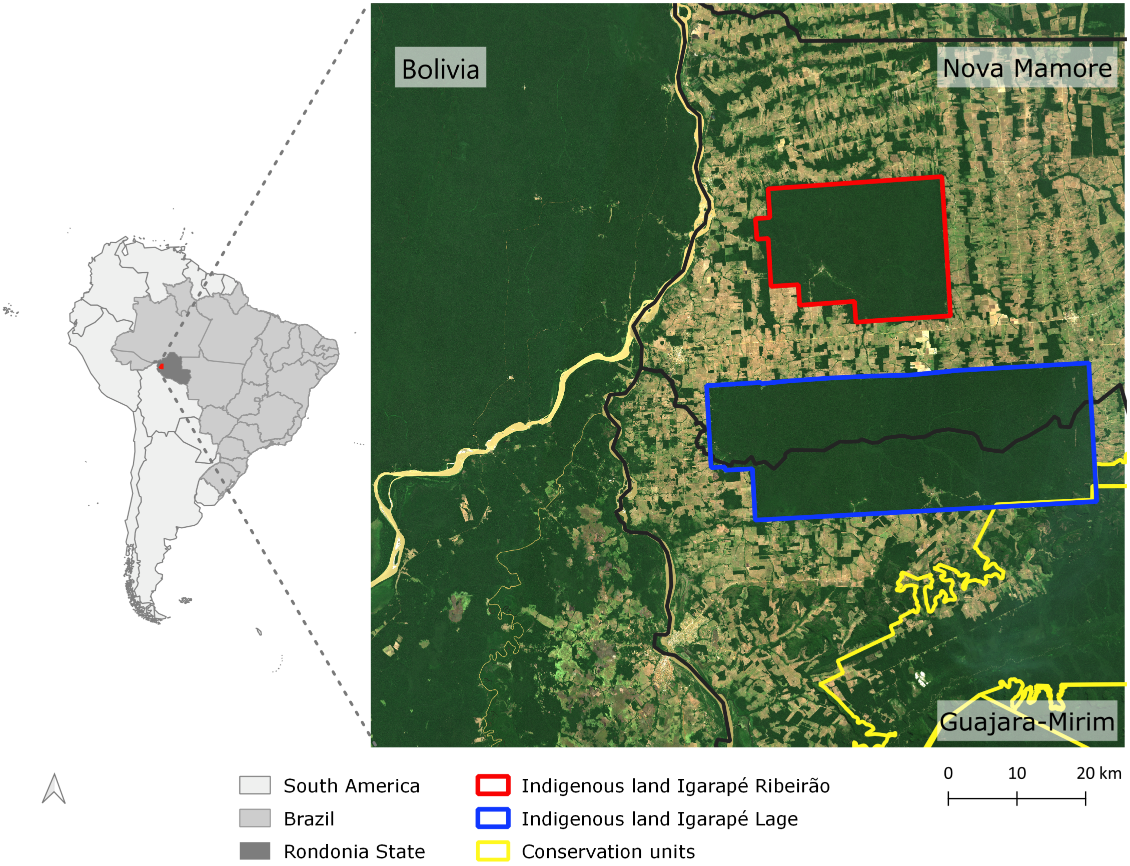
Our hypothesis is that, by using time series of Earth Observation imagery, we can improve the spatial accuracy of deforestation maps. We achieve this by means of high-dimensional data cubes using all available images coupled to machine learning algorithms. This is a consistent and efficient way of mapping deforestation over large datasets of EO imagery. In this paper, we introduce a novel method that uses time series build from a Sentinel-2A data cube to map annual deforestation over the Amazon forest using the full depth of available EO data. This method can improve deforestation detection in the highly dynamic and cloud contaminated time series of tropical forest (MUELLER et al., 2016). Besides, our approach is automated and can be scaled up to the whole Amazon biome, making it suitable to be used in deforestation monitoring systems.

4.3 Materials and methods

4.3.1 Study area

Our study area is located in the Amazon rainforest between Brazil and Bolivia. It displays a stark contrast between developed anthropic activities and natural forests with different levels of deforestation (see Figure 4.1). On the Brazilian side is the state of Rondônia, which has an area of 237.7 thousand km^2. Between 1989 and 2019, it had lost 11.1 thousand km^2 of forest (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, 2020). The amount of cattle in Rondônia has increased from 1.59 to 14.36 millions (an increase of $\approx$ 800%), from 1989 to 2018 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2018), being the major deforestation driver in this State. Our study area also includes two indigenous lands: *Igarapé Ribeirão*, in the municipality of Nova Mamoré, and *Igarapé Lage*, in the municipality of Guajará-Mirim. Additionally, there are three land reserves: *Rio Ouro Preto Extractive Reserve*, *Rio Ouro Preto Biological Reserve*, and *Rio Pacaás Novos Extractive Reserve*. On the other hand, the Bolivian side of our study area includes a state park for environmental protection in the province of *Federico Román*, in the department of *Pando*.

Figure 4.1 - Left: location of the study area. Right: Sentinel-2 image from 28 September, 2019 (bands 432, Tile 20LKP), indigenous lands, and conservation units of the study area.



4.3.2 Data

We used 36 images of Multispectral Instrument (MSI) on-board Sentinel-2A located in the tile 20LKP, from August 2018 to July 2019. These images were processed to surface reflectance (level 2A) using the software Sen2Cor version 2.8. In our data preprocessing we observed that, on average, half of each time series ($\approx$ 19 observations) is clouded.

We applied the Fmask4 cloud detection algorithm to the images in our study area following the results on (SANCHEZ et al., 2020b), which compares MAJA, Sen2Cor, s2cloudless, and Fmask4, and points to the better performance of the latter algorithm over the Amazon forest on Sentinel-2 images. To fill in the no-data values introduced by Fmask4, we applied a linear interpolation along the time dimension.

We resampled the Sentinel-2A bands of 20 m of spatial resolution to 10 m using bilinear interpolation. Besides the MSI bands, we used three spectral indices: Normalized Difference Vegetation Index (NDVI) (HUETE et al., 1985), Enhanced Vegetation Index (EVI), Normalized Difference Moisture Index (NDMI) (CIBULA et al., 1992). The attributes are listed in Table 4.1.

Table 4.1 - Spatial and spectral resolution of Sentinel-2.

Attribute	Resolution (m)	Wavelength (nm)
B02 - Blue	10	490
B03 - Green	10	560
B04 - Red	10	665
B08 - Broad infrared (bnir)	10	842
B8A - Narrow infrared (nnir)	20*	865
B11 - Short-wavelength infrared (swir1)	20*	1610
B12 - Short-wavelength infrared (swir2)	20*	2190
Enhanced Vegetation Index (EVI)	10	-
Normalized Difference Vegetation Index (NDVI)	10	-
Normalized Difference Moisture Index (NDMI)	10**	-

* Attributes later resampled to 10 meters. A
** Using resampled attributes.

SOURCE: Gascon et al. (2017)

.

Vegetation indices have been widely used to monitor vegetation conditions, such as biomass production, plant health, plant stress, water use, etc. Some vegetation indices most commonly used are NDVI and EVI, both been used used to quantify vegetation greenness, however EVI is more sensitive to differences in densely vegetated areas (WEIER; HERRING, 2000). NDMI uses the shortwave infrared band (swir), which is sensitive to changes in the water content of vegetation canopies, hence is better than NDVI to map clearcut and partial harvests in forest areas (WILSON; SADER, 2002; SCHULTZ et al., 2016).

We organized the observations in chronological order, creating a data cube with four dimensions (longitude, latitude, time, and attributes) (APPEL; PEBESMA, 2019). These data were generated and stored in the scope of Brazil Data Cube project[4].

[4] For more details, see http://brazildatacube.org.

Our training data set has 481 samples distributed into three classes: Forest (144), Other (295), and Deforestation (42). These samples were collected through visual interpretation from Sentinel-2A images and additionally assisted by high spatial resolution imagery available at Google Earth, from August 2018 to July 2019. The time series corresponding to the samples were extracted from our Sentinel-2A data cube.

4.3.3 Classification

We selected the two best models after training different combinations of attributes listed in Table 4.1, using the training data set. To accomplish this, we ran a set of 33 k-fold experiments ($k = 10$) using a Random Forest (RF) model with a thousand trees. The criterion used for selection was the highest median of F_1 score value for the Deforestation class.

We used Random Forest to build our classification models of deforestation. RF is one of the most used machine learning models applied to the classification remote sensing data (BELGIU; DRAGUT, 2016). It explores the solutions to classification problems by randomly and recursively building decision trees of observations and variables. RF uses an ensemble of decision trees to classify unlabeled inputs by means of a majority voting schema (BREIMAN, 2001).

We generated two maps of deforestation using the two aforementioned RF models, which were trained using the time series of the 481 sample data set. During training, we used RF models of 1000 threes and the full depth of Sentinel-2A time series, comprising 36 observations over time. The parameter for choosing the best forest RF models was the GINI index (BREIMAN, 2001).

Finally, we applied a spatial Bayesian smoothing algorithm using a 3 × 3 window. This method reclassifies the pixels based on the RF probabilities associated to each class and each pixel. The algorithm changes those pixels classes with high entropy to the neighborhood's class with low entropy using Bayesian inference. In these steps, we extensively used sits (CAMARA et al., 2018), an open source software package developed by our research team for the *R* environment for statistical computing and graphics (IHAKA; GENTLEMAN, 1996). Finally, the classes with higher probability were chosen for each pixel.

4.3.4 Accuracy assessment

We assessed the accuracy of our classification maps following the good practice guidelines by (OLOFSSON et al., 2014). The sampling and validation were done separately for both classified maps. We selected strata based on the area of each class in our classification maps. For both maps, we assumed a target standard error for the overall accuracy of ≈ 0.025.

To conjecture an accuracy, we set up a bootstrap experiment on which we selected 42 random samples for each class (38 for training, 4 for test); we selected this number because it matches the number of training Deforestation samples. Every ten iterations, we merged the test results and computed its user accuracy (UA). We repeated this entire process 100 times and used the median as the conjectured accuracy to determine the number of validation samples, which are 252.

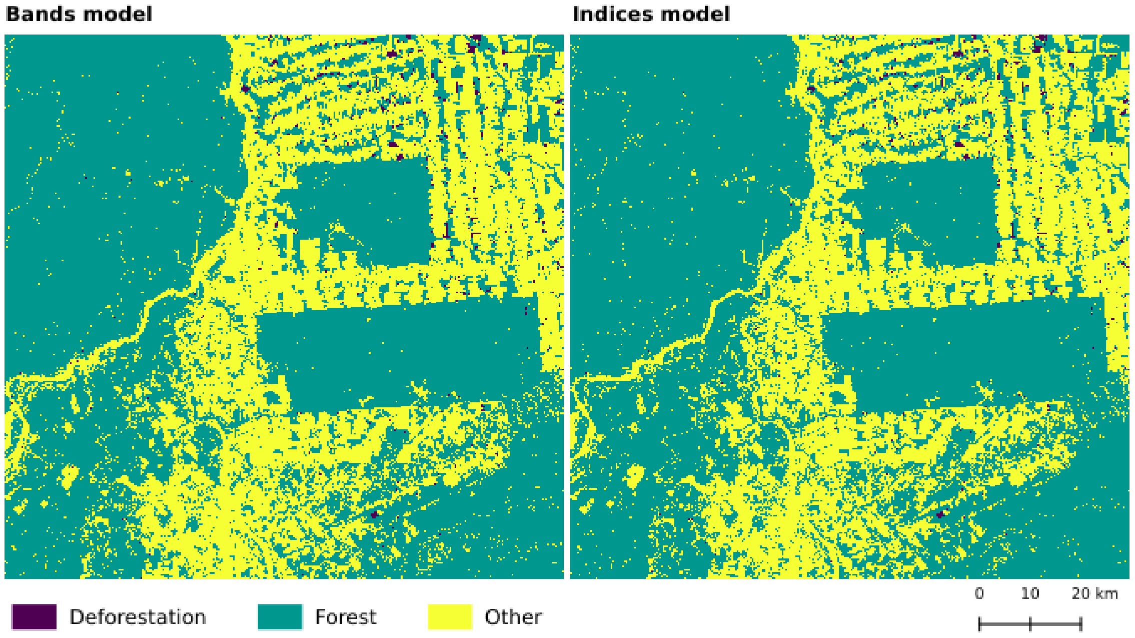
Then, we collected two independent sets of 252 samples (one for each map) following an equal allocation sampling design: 84 for Forest, 84 for Other, and 84 for Deforestation. This validation data set was collected through visual interpretation, the same way as the training data set. Finally, we calculated the confusion matrix, producer's and user's accuracy, bias-corrected error estimates, which are presented in the next section.

4.4 Results

The two models that presented the highest median of F_1 score for the Deforestation class were: 1) the combination of the blue, bnir, green, nnir, red, swir1, and swir2 bands (hereafter *Bands*); and 2) the combination of EVI, NDMI, and NDVI indices (hereafter *Indices*). For the *Bands* model, the median of F_1 score for Deforestation is 96.9%. For the *Indices* model, the median of the F_1 score for Deforestation is 95.6%.

Our results of the two RF classifications using Sentinel-2A data cube are shown in Figure 4.2. For the *Bands* classification, the mapped areas are 8,688 ha (0.7%) for Deforestation, 765,022 ha (63.5%) for Forest, and 431,894 ha (35.8%) for Other. For the *Indices* classification, the classes areas are 7,736 ha (0.6%) for Deforestation, 784,021 ha (65%) for Forest, and 413,848 ha (34.3%) for Other. By comparing *Bands* to *Indices*, we observe differences of 952 ha (+12.3%) in the class Deforestation, 18,999 ha (–2.4%) in Forest, and 18,046 ha (+4.4%) in Other.

Figure 4.2 - *Bands* and *Indices* classified maps using MSI/Sentinel-2A data cube from August 12th of 2018 to July 28th 2019.



For both classification maps, it is possible to verify visually that the deforestation in 2019 occurred mostly near Other areas. Both the *Igarapé Ribeirão* and the *Igarapé Lage* indigenous lands presented deforestation areas in 2019. However, in the *Igarapé Ribeirão* indigenous land, we observed deforestation occurring mainly near its borders.

Following the guidelines proposed by (OLOFSSON et al., 2014), we need to inform the conjectured UA for each class to calculate the size of the validation data set. To determine these parameters for both classification models, we setup a bootstrap experiment of 100 rounds.

Based on the bootstrap results for the *Bands* model, we conjectured an user accuracy (UA) of 82% for Forest, 81% for Other, and 84% for Deforestation. For *Indices* model, we conjectured 81% for Forest, 81% for Other, and 82% for Deforestation. Both conjectures resulted in a sample size of 252.

To validate the maps, we decided to allocate equally the number of samples for each class (84 samples). These samples were randomly and independently selected from the class strata of each map. We used two validation data sets, one for each classification map. The confusion matrices are shown in Tables 4.2 and 4.3.

Table 4.2 - Confusion matrix of the *Bands* classification model.

		Predicted			
		Deforestation	Forest	Other	Producer accuracy
Reference	Deforestation	75	0	9	89.2%
	Forest	1	76	7	90.4%
	Other	1	18	65	77.3%
	User accuracy	97.4%	80.8%	80.2%	
	F_1 score	93.1%	85.3%	78.7%	

Table 4.3 - Confusion matrix of the *Indices* classification model.

		Predicted			
		Deforestation	Forest	Other	Producer accuracy
Reference	Deforestation	74	1	9	88.0%
	Forest	1	70	13	83.3%
	Other	2	8	74	88.0%
	User accuracy	96.1%	88.6%	77.0%	
	F_1 score	91.9%	85.8%	82.2%	

In the *Bands* classification, we obtained an UA of 97.4% for the Deforestation class, surpassing our conjecture. However, we fell short in the case of the Other class, where we expected 81% UA, but we obtained 80.2%. For the Forest class, we get an UA of 80.8%, almost equal to our conjecture. In the *Indices* classification, the Deforestation UA is 96.1%, greater than our conjecture. For the Other class, we obtained an UA less than our conjecture. Finally, the Forest UA was 88.6%, exceeding our conjecture.

For Deforestation, the *Bands* classification has highest values of F_1 score when compared with *Indices* classification. The Forest and Other classes had better results of F_1 score using *Indices* than using *Bands* model. Our classifications have overall accuracy of 88.9% for *Bands* and 84.9%for *Indices*.

4.5 Discussion

The MSI sensor on board of the Sentinel-2A satellite has 13 bands from which can be derived more than 240 spectral indices[5]. These attributes have different discriminatory power in regards of deforestation. In consequence, we used our training samples to test different attribute combinations and to verify their accuracy and robusticity for classification. The k-folds technique guided us in the process of model selection, allowing us to compare different attribute combinations (LEVER et al., 2016). From a list of 11 bands and three indices, we ran 33 k-fold experiments for each combination.

We selected the two best models, one using the bands blue, broad nir, green, narrow nir, red, swir1, and swir2 bands (*Bands*), and the other using the indices EVI, NDMI, and NDVI (*Indices*). Several studies have successfully tested and applied indices to detect deforestation (GRINAND et al., 2013; HAMUNYELA et al., 2016; SCHULTZ et al., 2016; ADARME et al., 2020). Also, other authors obtained robust results using spectral mixture models, where the most common endmembers are green vegetation, non-photosynthetic vegetation, and bare soil (ASNER, 2009; SOUZA JUNIOR et al., 2013; SCHULTZ et al., 2016). However, to the best of our knowledge, there is not in the literature any method relying exclusively on spectral bands for detecting deforestation.

The use of spectral bands has many advantages because they convey specific information for different land features. For example: to estimate chlorophyll content, the best wavelengths regions — or bands — of the electromagnetic spectrum are the green and red (GITELSON; MERZLYAK, 1994; CLEVERS; KOOISTRA, 2012). The red band is also used to classify soil. In the nir region, healthy plants have higher reflectance due to leave structures and it can be used to assess canopy variation in vegetation biomass. The water in vegetation is known to absorb energy in the infrared region (broad nir, narrow nir, swir 1 and swir 2), hence these bands can be used to classify this cover (JENSEN, 2009). This absorption is more dominant in the swir region. Conversely, dry soil tends to exhibit a much higher reflectance in this same wavelength region (JENSEN, 2009), which can be useful for deforestation detection.

The two classification models applied in this study use the Random Forest algorithm. Both obtained similar accuracy using either seven spectral bands or three vegetation indexes. However, the accuracy of the classification using *Bands* is slightly and consistently better than the accuracy using *Indices*. Besides, the classification using *Bands* is smoother in appearance, with less salt and pepper effect (see Fig-

[5] Index DataBase - Sentinel-2A indices `https://www.indexdatabase.de/db/is.php?sensor_id=96`

ure 4.2). This could indicate that *Bands* model is delivering more information to the classification algorithm.

Figure 4.3 depicts some examples of the classification map. In the Figure 4.3a, it is possible to verify a small deforested area on the upper left that was detected by both models. In general, the classification using *Bands* classifies deforested areas homogeneously, while the model that uses *Indices* classifies deforestation patches heterogeneously. The *Bands* model classifies deforested patches with less discontinuities, similar to the maps produced by PRODES and GFW (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, 2020; HANSEN et al., 2013).

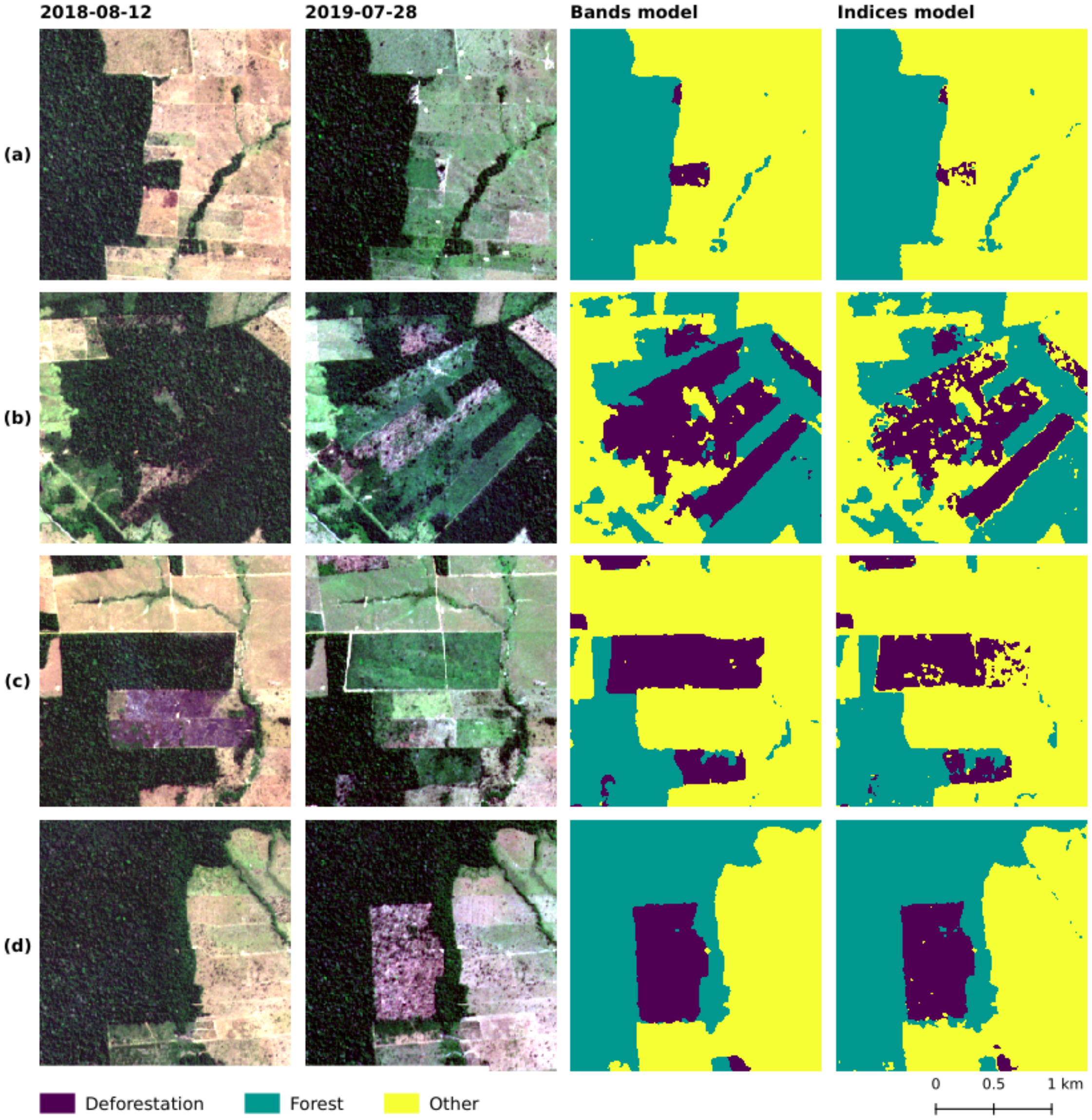
In Figure 4.3b, it is possible to observe that the classification using *Indices* identified several spots of the Other class within the deforestation patch. This may be an indication that, in the time series, these areas did not have a spectral responses similar to forest. However, it was not possible to verify visually. This can happen because of the vegetation indices, EVI, NDVI, and NDMI, which are related to plant biomass, may have captured something that the spectral bands, used separately, did not identify, e.g. the vegetative vigor. This can be an indication of forest degradation.

In Figure 4.3c, where there was a clear-cut, the classification using *Indices* did not detect the full extent of deforestation, while the *Bands* model map the entire deforested patch. Also, we can note that *Indices* model produces a border effect. The sensitivity of the *Indices* model in classifying small forests areas can be observed in Figure 4.3d, where the model classified a riparian forest in the upper left corner.

In the validation step, the Forest class had confusion mainly with the Other class, in both classifications. This happened because of the classification models were not trained to identify different percentage of tree cover, as done by (HANSEN et al., 2013). Also, it was verified that some areas of secondary vegetation had been mapped as Forest. This problem could be solved by exploring long-term time series.

The lowest accuracy occurred in the Other class (F_1 score equal to 78.8% for *Bands* and 82.2% for *Indices*). This class includes heterogeneous features such as pasture and abandoned and degraded areas, which were not considered in this study. Possibly by adding these features as new classes in the model might increase accuracy. The identification of these areas could be of interest to other types of analysis such as those of (GIBBS et al., 2015b), (PICOLI et al., 2020), and (PARENTE et al., 2017).

Figure 4.3 - Deforestation patches comparisons between *Bands* and *Indices* models. The two columns from left show Sentinel-2A images of start date and end date (true color). The central coordinates of the patches are (in WGS84): (a) $(-64.97182 \ -10.28505)$; (b) -65.30635 -10.05332; (c) $-64.8961 \ -10.28891$; (d) (-64.96943 -10.33271).



In order to validate the maps, we followed the good practice guidelines proposed by (OLOFSSON et al., 2014). To calculate the sample size, the authors suggest that users conjecture the UA based on past experience with similar works. In this study, we propose a less subjective procedure, by using the training samples and bootstrap techniques to conjecture the UA for strata.

Since the UA estimates the probability that a pixel be classified correctly according to the reference, it can be used to check the reliability of the classification model. Our accuracy obtained with the validation was satisfactory for the detecting deforestation. The observed UA was 97.4% for the *Bands* model and 96.1% for the *Indices* model. Additionally, the PA was 89.2% for *Bands* and 88% for *Indices*, showing that both classifiers are robust for mapping deforestation.

These accuracies are promising when compared to some important works in the deforestation literature (Table 4.4). All these works use the space-first approach. For example, (SOUZA JUNIOR et al., 2013) use a decision tree to detect deforestation in Landsat 5 and 7 images, the results achieved were 85% for UA and 92% for PA. It is important to highlight that, contrary to our work, this classification included a manual step of editing deforestation polygons by specialists.

Another example is the work of (ADARME et al., 2020) with NDVI derived from Landsat 8 images and several machine learning methods based on Deep Learning and Support Vector Machine algorithms on the Brazilian Amazon forest. They obtained a F_1 score for deforestation ranging from 39% to 63% (depending on the number of training areas), while our time series approach reaches F_1 score slightly above 91% (see Tables 4.2 and 4.3). (SCHULTZ et al., 2016) used Landsat and BFAST on time series of 8 vegetation indices, obtaining satisfactory accuracies above 80% in the best case scenario (NDFI). Their results are on the same range as ours and both support our claim regarding the use of time series for deforestation monitoring of the Amazon forest.

When we comparing our proposed method with previous works in the literature (Table 4.4), we observed that, to the best of our knowledge, our approach is the only one that combines time series and machine learning algorithm. Among several deforestation detection studies observed in the literature, the use of Random Forest, Deep Learning, and Support Vector Machine is a applied to multitemporal images, such as the works by (GRINAND et al., 2013) and (ADARME et al., 2020). Regarding time series and deforestation detection, the most used method is BFAST (HAMUNYELA et al., 2016; SCHULTZ et al., 2016), a algorithm based on the statistical additive decomposition of time series. This method is generally applied to a single attribute time series.

Table 4.4 - Summary of selected approaches for deforestation mapping. We report the best accuracy obtained by each work.

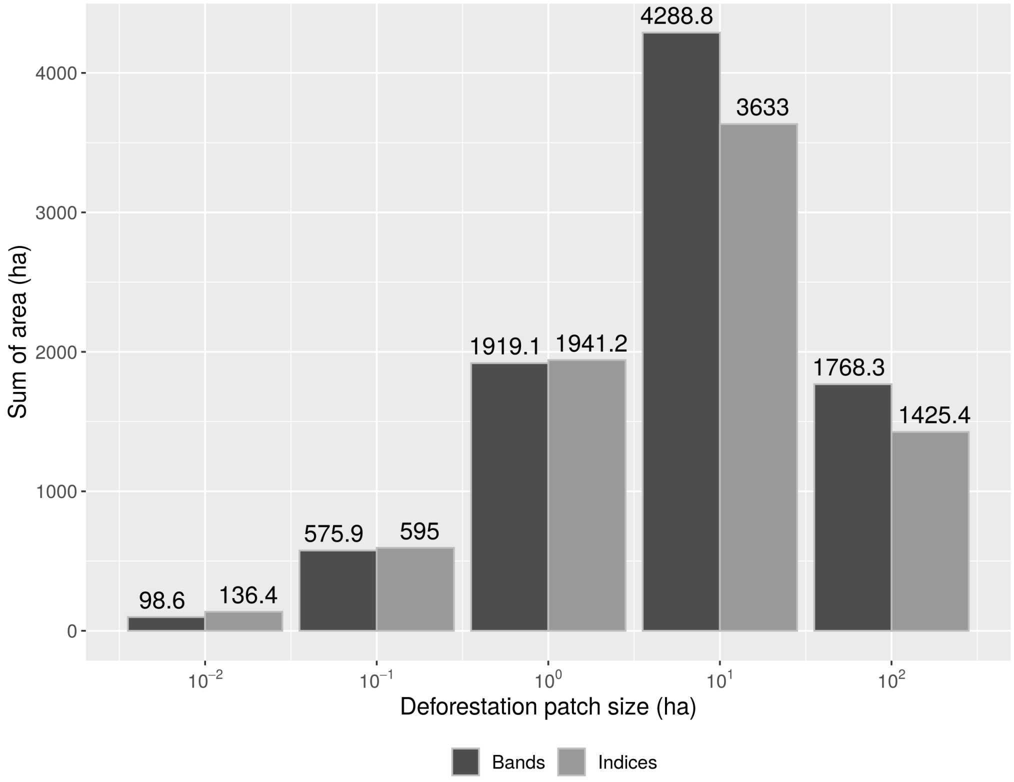
Method	Description	Accuracy
Random Forest using Landsat 7 (GRINAND et al., 2013).	They classified deforestation in some regions of Madagascar using Random Forest using Landsat B1, B4, B5, NIRI, and NDVI from 2000 to 2010.	60.7% (UA); 41.0% (PA); 48.9% (F_1 score*)
Decision tree and manual editing using Landsat 5 and 7 (SOUZA JUNIOR et al., 2013).	They combine spectral mixture model (green vegetation, non-photosynthetic vegetation, and soil components) with decision trees. The method comprehends both manual and automatic steps. The work was applied in the Brazilian tropical rainforest area.	85% (UA); 92% (PA); 88.3% (F_1 score*)
Breaks For Additive Season and Trend (BFAST) Monitor method using Landsat 5 and 7 time series (SCHULTZ et al., 2016).	The performance of Landsat time series (LTS) of eight vegetation indices was assessed for monitoring deforestation across the tropics. Three sites (Brazil, Ethiopia, Vietnam) were selected based on differing remote sensing observation frequencies, deforestation drivers and environmental factors. For the Brazilian area the best result to deforestation detection was achieved using the NDFI index.	80.4 (UA*); 89.3 (PA*); 84.6% (F_1 score*)
Breaks For Additive Season and Trend (BFAST) using Landsat 5 and 7 time series (HAMUNYELA et al., 2016).	They used the NDVI derived by Landsat 5 and 7 (from April 1984 to December 2014) to detect deforestation events applying BFAST, in one area in Brazil and other in Bolivia tropical forests. To reduce the spatial seasonal variations they propose to calculate the spatially normalized NDVI (sNDVI). Thet detected deforestation from sNDVI time series using BFAST. The best result, to deforested detection, was achieved using sNDVI time series calculated applying using a spatial context to account for seasonality.	Brazil: 93.9% (UA*); 94.9% (PA*); 94.4% (F_1 score*) Bolivia: 98.4% (UA*); 97.4% (PA*); 97.9% (F_1 score*)
Deep Leaning and Support Vector Machine algorithms using Landsat 8 multitemporal images (ADARME et al., 2020).	They classified the deforestation in two Brazilian regions, one in the Amazon and other in the Cerrado biome. For this they used the NDVI by Landsat 8 images acquired at tow different dates, and evaluates the methods to detect deforestation: Early Fusion, Siamese Network, Convolutional Support Vector Machine, and Support Vector Machine. The images were divided into 15 tiles: four tiles were selected for training, two tiles for validation, and nine tiles for testing. The best result to deforestation detection in Amazon region was using four tiles for training and the Early Fusion algorithm.	63.2% (F_1 score)
Random Forest using Sentinel-2A time series (by the authors)	We introduce a method of using time series build from Sentinel-2A data cube to map deforestation over the Amazon forest. For more information see Section 2. The best result to deforestation detection was achieved using the bands combination blue, bnir, green, nnir, red, swir1, and swir2.	97.4% (UA); 89.2% (PA); 93.1% (F_1 score)

* Values estimated from accuracy figures in the original papers.

As noted by (RICHARDS et al., 2017; KALAMANDEEN et al., 2018; HANSEN et al., 2020), the size of the patches of deforestation has been reduced over the past few years. This trend poses a challenge for deforestation monitoring systems. In Figure 4.4, it is possible to observe that ≈ 30% of the deforestation areas mapped by the *Bands* model have 1 ha or less. This value goes up to ≈ 34% for *Indices*. Part of this increase can be explained by the salt and pepper effect discussed earlier most frequently present in the *Indices* model.

Although the minimum mapping unit is 0.09 ha for 30 m images, e.g. Landsat, the effective unit tends to be larger as it is hard to validate isolated pixels. On this scale, the spatial attributes of deforestation like shape and texture vanish. These spatial attributes are commonly used by experts to validate the deforestation. Furthermore, the lower spatial resolution the greater the edge effect, as there is a spectral mixing of targets in the pixels belonging to borders. The use of Sentinel-2A, which has a minimum mapping unit of 0.01 ha (10 m resolution), can be a feasible solution for small patches detection, once it has a smaller area of mixing targets.

Figure 4.4 - Distribution of deforestation patches in the *Bands* and *Indices* classifications.



Due to its high accuracy, we believe our classification method could have a positive impact by being integrated into existing non-automatic deforestation monitoring systems. Another application of our classification method could be supporting non-automatic deforestation monitoring systems, that usually use exclusion masks to avoid visiting already reviewed areas, to detect deforestation in secondary vegetation. However, our method is constrained by the the inability to distinguish among natural and anthropogenic deforestation and we avoided the identification of forest degradation such as selective cut or fire. We also acknowledge that time series classification misses important spatial attributes such as shape, size, shadow, tone, texture, and location.

4.6 Conclusions

Combining machine learning and time series improved the identification of deforestation in the Amazon rainforest. The successive and consistent observations of a place in time series capture the forest dynamics and deliver more information to classification algorithms than using independent multi-temporal analysis. Time series make better use of the data available, decreasing the chances of miss-classification due to noisy observations.

We investigated which combination of Sentinel-2A attributes performs better deforestation detection. Here, we generated and validated two classifications. The first one uses a Random Forest model with the bands blue, green, red, broad infrared, narrow infrared, short-wavelength infrared 1, and short-wavelength infrared 2. The second one also uses Random Forest model, but the inputs are the spectral indices Enhanced Vegetation Index, Normalized Difference Vegetation Index, and Normalized Difference Moisture Index. We conclude that, for deforestation detection purposes, the first model (using bands) performed better. In addition, the 10 m spatial resolution of Sentinel-2A images allows mapping smaller areas of deforestation that would not be detected in coarser spatial resolution sensors, such as OLI/Landsat 8 (30 m) and WFI/CBERS-4 (64 m). This could solve part of the challenge that, over the past few years, it has been observed that deforestation patches are becoming smaller (KALAMANDEEN et al., 2018).

By using the time series of EO imagery, we obtained reliable classification models (UA > 96%) with accuracy above those reported in the literature for detecting deforestation. This is possible due to the availability of high-dimensionality data cubes coupled with machine learning techniques.

The maps generated by the proposed method can support the enforcement of policies and agreements to fight deforestation, such as the Brazilian Forest Code and the Soy and Cattle Moratorium. Without reliable deforestation maps, the monitoring and law enforcement activities lose their effect on supporting environmental policies and agreements. Furthermore, the maps we produced can also feed climate change as well as land use models.

Our approach demands low expert intervention, and hence can easily be automated and scaled. Because of these characteristics, it is feasible to integrate our method in deforestation monitoring systems. Our method was facilitated by a data cube platform specialized in satellite image time series, the Brazil Data Cube platform. However, the application of this method in large areas will require significant computational infrastructure. Thus, building institutional mechanisms, involving public, private, and multilateral organizations, is required to guarantee continuity of investments in wide-ranging monitoring systems.

4.7 Code and data availability

The code used in our experiments is provided under the GNU General Public License v3.0 and is available in (SANCHEZ, 2020). The R package *sits* (Satellite Image Time Series) provides tools for handling time series of Remote Sensing images, including data retrieval, visualization, and machine learning methods (e.g. Random Forest). The latest release of *sits* is available on GitHub at https://github.com/e-sensing/sits.

Our results are available in the PANGAEA platform (SANCHEZ et al., 2020a). They contain the deforestation maps of 2019 in GeoTIFF format at 10 meter resolution as well as the files with the training data set (481 samples), validation data sets (252 samples for each *Bands* and *Indices*) in CSV format.

5 REPRODUCIBLE GEOSPATIAL DATA SCIENCE: EXPLORATORY DATA ANALYSIS USING COLLABORATIVE ANALYSIS ENVIRONMENTS[1]

5.1 Abstract

The answers to planetary problems could be hidden in gigabytes of satellite imagery from the last 40 years. Unfortunately, scientists lack the means for processing such amount of data as they are used to work over small quantities of satellite images. To amend this issue, we propose the use of web services from Big Earth data platforms along collaborative analysis environments. Both Web services and collaborative analysis environments fit the hypothesis-test workflow followed by researchers while writing analysis routines. Besides, the early use of Big Earth data structures eases the subsequent process of scaling analysis up to larger extensions. To test our proposal, we use our own Big Earth observation data platform, on which decades of satellite images are arranged into data cubes. By using our Web services platform, we integrate those data cubes into our collaborative analysis environment (a Jupyter notebook). Since our analysis routines consume the same data structure of the whole data sets, it is easier to scale up the analysis.

5.2 Introduction

The process of analyzing Earth observation data is a combination of science and art. It requires knowledge, perseverance and some resignation for the effort put on failed tests which never reach the final publications. To advance their research, scientists rely on a hypothesis-test cycle and diaries —or notebooks— to keep record of their findings. This process also relies on computer code, which scientists write their own way, following the same hypothesis-test cycle over small data sets. Nowadays, computers also help scientists to manage their digital notebooks based on concepts such as Literate Programming and Overlay Journals. Furthermore, these notebooks are being taken to the web in the form collaborative analysis environments, which are on-line documents that mix code, data, descriptions, and tables to summarize the results of scientific research. This electronic approach to analysis fits well the current data distribution

[1] This chapter is an adapted version of the paper on *Revista Brasileira de Cartografia*: Sánchez, A., Vinhas, L., Queiroz, G., Simoes, R., Gomes, V., Assis, L.F., Llapa, E., Camara, G., 2018. Reproducible geospatial data science: Exploratory data analysis using collaborative analysis environments. Rev. Bras. Cartogr. 70, 1844–1859. It is also an extended version of the paper presented in the XVII Brazilian Symposium on GeoInformatics (GEOINFO 2017): Sanchez, A., Picoli, M., Andrade, P. R., Simões, R., Santos, L., Chaves, M., Begotti, R., & Camara, G. (2019). Land Cover Classifications of Clear-cut Deforestation Using Deep Learning. Geoinfo, 48–56.

model based on files (KNUTH, 1984; HEY et al., 2009; PEREZ; GRANGER, 2007).

However, this approach is unsuitable to the analysis of large regions of space and time; as a result, there are few global scale analyses in the scientific literature. Besides, a file-based model —as the one used to distribute satellite imagery— fosters problems such as data duplication and lack of traceability. On the other hand, global data sets are either unavailable or just too large for independent result validation. Both scenarios worsen the current scientific reproducibility crisis (BAKER, 2016; ANONYMOUS, 2016).

This situation shows the issues of scaling up software routines for data analysis. Putting aside those related to computing power —they are already addressed in the literature on the data deluge or big data— we focus on the transit from small to large datasets. It is important for scientists to keep fast and short iterations of think-code-test and to minimize the amount of re-work incurred while scaling up analysis (BELL et al., 2009; BOYD; CRAWFORD, 2012; LI et al., 2016).

We addressed this problem by setting up collaborative analysis environments along big Earth data web services. The former enables fast iterations of the hypothesis-test cycle while the latter enables scientist to analyze increasingly larger data sets. In other words, the earlier scientist use with Big Earth observation data structures, the easier to scale analysis to larger extensions.

In this paper, we examine how Web services provided by big data platforms can be integrated into the analysis workflow of Earth observation data. To achieve this, we briefly introduce a computing platform – developed by us— and its web services (Section 5.3 and Section 5.4). Then, we describe analysis environments and how the into the scientists' workflow (Section 5.5). Finally, we test our approach by setting up Jupyter notebook – a collaborative analysis environment – in which we mixture the web services provided by our platform and the analysis analytical tools provided by the Python programming language.

5.3 The e-sensing platform

The Brazilian National Institute for Space Research (INPE) runs the e-sensing project. This project is building a platform for scientist to research Land Use and Land Cover Change (LUCC). The platform sorts decades of satellite images into multidimensional space-time arrays.

The main requirements to these platforms are analytical scaling, software reuse, collaborative work, and replication. Analytical scaling is about moving data and code among computing platforms with little or no modifications at all. Software reuse refers to the ability to run code from different origins. Collaborative work and replication are about sharing and replicating analysis results. We address software reuse, collaborative work, and replication by using open source and open access software and data. Our platform only hosts open source software and open access data such as MODIS and LANDSAT images (CAMARA et al., 2016a; STONEBRAKER et al., 2009).

We have been using our platform to classify time series of vegetation indexes of the Amazon and Cerrado biomes into LUCC classes. Later, during post-processing stages, we analyze the LUCC trajectories over time. But the data workflow inside our platform relies on a mixture of technologies such as scripting languages (*R*, Python, Bash), distributed storage (SciDB, Hadoop), and operating system tools. As a result, the scientific reproducibility of our results is compromised. Therefore, we chose web services as the way to expose our platform computing capabilities while hiding its internal complexities (ASSIS et al., 2016; CAMARA et al., 2016b; LU et al., 2016; MACIEL et al., 2019; MAUS et al., 2016).

On the other hand, the CEOS Data Cube Platform (CEOS-ODC) handles storing, accessing, and managing metadata of remotely sensed data. CEOS-ODC is built on top of the Australian Geoscience Data Cube. Just as e-sensing, the CEOS-ODC platform can process large amounts of satellite imagery using open source tools. However, they employ different analysis and architectures. While e-sensing is focused on time series analysis, CEOS-ODC puts spatial before temporal analysis. Regarding architectures, e-sensing is built on top of array databases while CEOS-ODC is built around the programming language python and data files; this difference is subtle but important since databases are independent of programming languages. As a consequence, the e-sensing platform is able to run analysis written in different languages while CEOS-ODC is constrained to python scripts (CEOS, 2016).

5.4 A Web Service for retrieving time series

Sharing and re-using computer resources has been important since the 90s because writing software is error-prone and high performance hardware is expensive. Nowadays, Web services are a common way to address this matter. Web services are the standardized way to access software and data over the World Wide Web independently of operating systems and programming languages. Through them, scientists can access the data and algorithms available in our platform. At the same time, web services hide complexities – such as mixed technologies, and distributed storage –

behind a uniform interface.

The Web Time Series Service (WTSS) retrieves time series of Earth Observation data for specific locations on Earth. WTSS reduces the gap between data and remote-sensing time-series clients through simple text representations using JSON (a standard file format). Traditionally, assembling time series of Earth Observation imagery is a time-consuming task because users need to sequentially open several image files, extract some pixels, and then store them. Instead, WTSS connects to a multidimensional array database and makes temporal queries on behalf of the client. WTSS exposes three main operations *list_coverages*, *describe_coverage*, and *time_series*. *list_coverages* returns a JSON list of the available coverages in the service. *describe_coverage* retrieves metadata of a specific coverage. Finally, the *time_series* operation retrieves specific time series. WTSS implementation is publicly available on-line (VINHAS et al., 2017).

Moreover, WTSS has clients for the QGIS software and for the scripting languages R and Python. These WTSS clients enable scientists to access our data from on-line analysis environments.

5.5 Interactive and collaborative analysis environments

Literate programming is a style of coding software in which programs are treated as pieces of literature. That is, natural and machine languages are weaved together into a document where thought order prevails over code optimizations. Its goal is to create programs easier to understand and maintain and to achieve this, literate programming makes explicit the reasoning behind the code (KNUTH, 1984).

Note how literate programming fits the way scientists analyses their data. Once data is collected, scientists make research questions, and then formulate hypotheses for later testing them on the data. The question making and hypothesis formulating is better described using natural language while data processing and hypothesis testing are automated using code.

The modern realization of literate programming is the on-line analysis environments. Using modern technologies, they add collaboration and interactivity to the traditional scientific notebooks and laboratory journals. Some examples are the *R* and Jupyter notebooks. It is worth noticing that *R* notebooks are focused in *R* while Jupyter notebooks support various programming languages. For this reason, we preferred the latter.

Statistical data analysis is crucial to science. From the computing perspective, the most popular and powerful computing tools for statistical analysis are *R* and Python. *R* is a computing environment designed for statistical analysis while Python is a general purpose programming language focused on readability and extensibility. Both support numerical processing, statistical data structures; the former natively while the latter trough code libraries such as SciPy. Both *R* and Python are supported by large communities of users coming from either the field of statistics or computer science. In this paper we preferred python because most of the authors come from computer science field (IHAKA; GENTLEMAN, 1996; VIRTANEN et al., 2020; REDMONK. . . ,).

IPython adds facilities to Python for scientific computing. IPython has an interactive command with tailor-made features for scientists, such as code completion, plotting, and parallel and distributed processing. These characteristics are taken to the web in the form of Jupyter notebooks. For example, the data and algorithms regarding the recent astronomic discovery of gravitational waves are available as Jupyter notebooks (KLUYVER et al., 2016; CANTON et al., 2014; USMAN et al., 2016).

5.6 Analysis of time series of vegetation indexes

Vegetation indexes are simple estimates of vegetation activity derived from satellite imagery. They are independent of measurement units and for this reason they are well suited for Land cover identification. However, satellite imagery is subject to noise which induces variance on the time series of vegetation indexes (HUETE et al., 1985; JIANG et al., 2008). Statistical analysis provides several tools for time series analysis; some of them are of common usage for image analysis (e.g. line fitting, Fourier decomposition, Whitaker smoother, and the Kalman filter) and classification (e.g. Dynamic Time Warping), particularly for noise removal and classification (ATKINSON et al., 2012). In this section we provide a trivial summary of analysis techniques because a complete discussion is beyond the scope of this paper.

Line fitting is the process of finding the straight line which minimizes the differences to the points in the time series. Line fitting is useful to find global trends in the data and it is the starting point for more complex fittings.

Fourier decomposition is a smoothing technique which is based on the Discrete Fourier Transform (DFT) and its inverse function. Assuming that time series are originally defined in the time domain, DFT converts time series data to the frequency domain while the inverse DFT convert from back from the frequency to the time domain. In the frequency domain, a time series is the sum of sinusoids characterized by a frequency. Higher frequencies correspond to noise. Smoothing is achieved by removing these high-frequency sinusoids and then reconstructing the time series using the inverse DFT (HEIDEMAN et al., 1984; JAKUBAUSKAS. et al., 2001).

The Whitaker smoother computes smoothed values for each observation using least squares over the linear combination of nearest observations, while penalizing the roughness of the smoothed results (ATZBERGER; EILERS, 2011; EILERS, 2003). The Kalman filter is an algorithm for estimating an unobserved quantity from a set of noise observations. As new observations are available, the Kalman filter improves its estimation, and due to its simplicity and speed, it is suitable for applications in engineering, econometrics, and more recently, remote sensing (GREWAL; ANDREWS, 2010; KLEYNHANS et al., 2011).

Dynamic Time Warping (DTW) is an algorithm that computes a similarity measure – a distance – between two time series. Given a set of time series of known land coverages (the patterns), we compute the DTW distances to a time series of an unknown land cover (the samples). The samples are assigned to the labels of the patterns with the shortest DTW distance (BERNDT; CLIFFORD, 1994). These analysis methods are applied to time series of vegetation indexes in the following section.

5.7 A collaborative environment - Jupyter notebook

We setup a Jupyter notebook for the exploratory analysis of time series of vegetation indexes. It mixes the web services provided by our platform and the analytical tools provided by the Python programming language. This notebook presents three common jobs regarding time series of vegetation indexes: Exploratory analysis, filtering or smoothing, and classification.

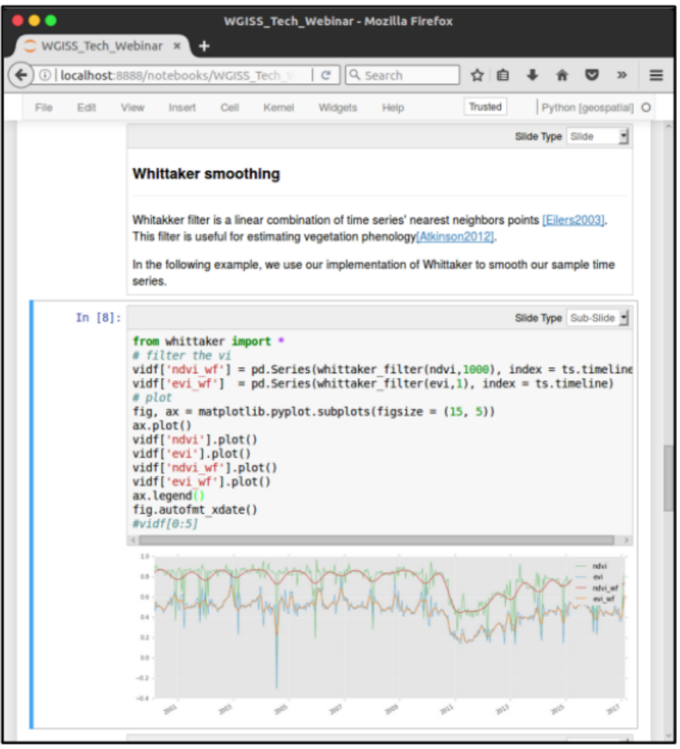
In the exploratory analysis, we get the data and then plot the time series and its location on a map. Listing 5.1 shows how to retrieve MODIS data into a data frame which is a table-like data structure. Lines 1 to 3 load the existing libraries, while lines 4 to 6 establish a point on Earth, some vegetation indexes, and where to find the Web Service. Line 7 retrieves time series from the Web Service, and finally, lines 9 to 13 arrange the data into a data structure called *data frame*.

Listing 5.1 - Get a time series into a Python pandas data frame.

```
import pandas as pd
from wtss import wtss
from tsmap import *
w = wtss("http://www.dpi.inpe.br/tws")
latitude = -14.919100049
longitude = -59.11781088
ts = w.time_series("mod13q1_512", ("ndvi", "evi"), \
     latitude, longitude)
ndvi = pd.Series(ts["ndvi"], index = ts.timeline) * \
       cv_scheme['attributes']['ndvi']['scale_factor']
evi = pd.Series(ts["evi"], index = ts.timeline) * \
      cv_scheme['attributes']['evi']['scale_factor']
vidf = pd.DataFrame({'ndvi': ndvi, 'evi': evi})
```

Once the time series is formatted as a data frame, it is possible to apply on it functions that receive and return data frame's columns as parameters. In this way, we smoothed our time series using the Whittaker smoother (Figure 5.1), the Kalman filter, and the Fourier decomposition.

Figure 5.1 - An on-line analysis environment for time series of Earth observation data. This environment displays a description of the Whitaker smoother, its Python implementation, and its results when applied to a time series of vegetation indexes.



Whittaker smoothing

Whitakker filter is a linear combination of time series' nearest neighbors points [Eilers2003]. This filter is useful for estimating vegetation phenology[Atkinson2012].

In the following example, we use our implementation of Whittaker to smooth our sample time series.

In [8]:

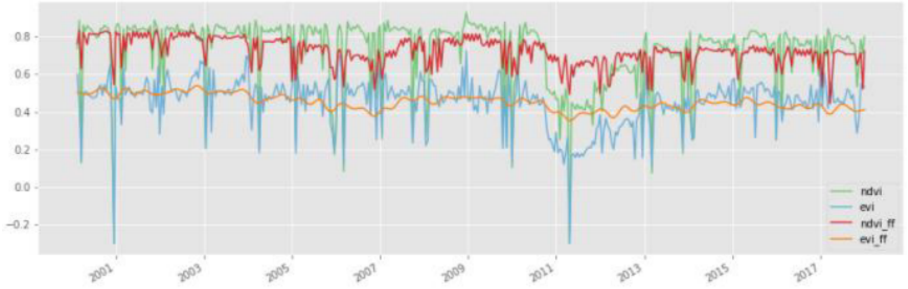
```
from whittaker import *
# filter the vi
vidf['ndvi_wf'] = pd.Series(whittaker_filter(ndvi,1000), index = ts.timeline
vidf['evi_wf']  = pd.Series(whittaker_filter(evi,1), index = ts.timeline)
# plot
fig, ax = matplotlib.pyplot.subplots(figsize = (15, 5))
ax.plot()
vidf['ndvi'].plot()
vidf['evi'].plot()
vidf['ndvi_wf'].plot()
vidf['evi_wf'].plot()
ax.legend()
fig.autofmt_xdate()
#vidf[0:5]
```

Listing 5.2 - Filter a time series using the Whitaker smoother.

```
1  from whittaker import *
2  vidf[’ndvi_wf’] = pd.Series(whittaker_filter(ndvi,1000), \
3                              index = ts.timeline)
4  vidf[’evi_wf’] = pd.Series(whittaker_filter(evi,1), \
5                             index = ts.timeline)
6  fig, ax = matplotlib.pyplot.subplots(figsize = (15, 5))
7  ax.plot()
8  vidf[’ndvi’].plot()
9  vidf[’evi’].plot()
10 vidf[’ndvi_wf’].plot()
11 vidf[’evi_wf’].plot()
12 ax.legend()
13 fig.autofmt_xdate()
```

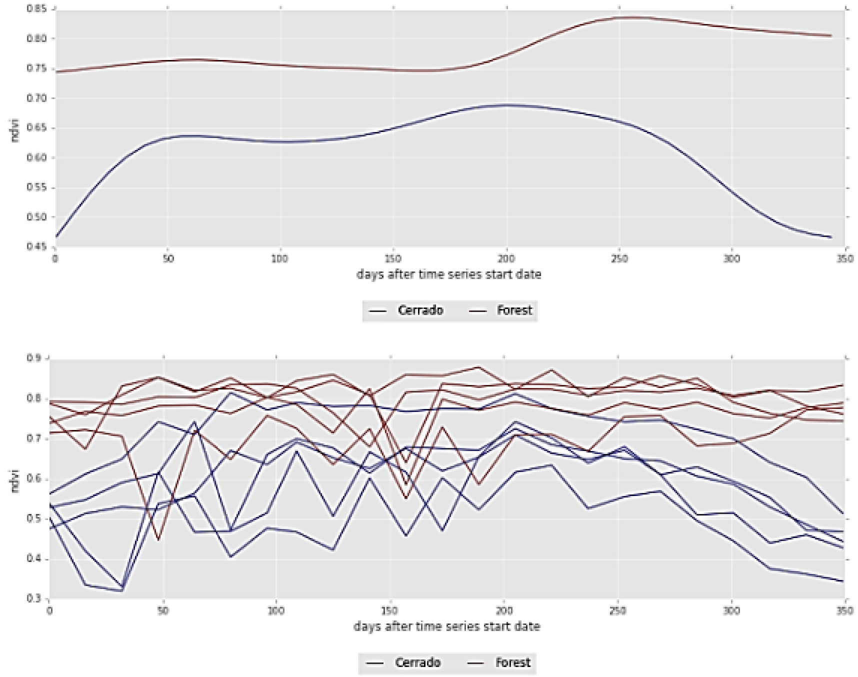
The code used to apply filters on the data is illustrated in Listing 5.2. Line 1 imports the filter which is applied to vegetation indexes (lines 2 and 4). The remaining lines of code print the filtered vegetation indexes along with the original data (lines 6 to 13). This code pattern is repeated for applying the Kalman filter and the Fourier decomposition (Figure 5.2).

Figure 5.2 - Fourier decomposition of time series of vegetation indexes.



The last example in our Jupyter notebook is classification. We used Dynamic Time Warping (DTW) to classify time series of vegetation indexes. We prepared a set of pattern time series corresponding to the land covers cerrado and forest. We also collected a set of sample points from which we know the latitude, the longitude and the land cover over a specific time interval; then we retrieved the time series of these points using WTSS. Figure 5.3 shows the time series of both patterns and samples. Listing 5.3 shows the code required to read the prepared files, retrieve the time series and to do the classification: Lines 1 and 2 load libraries while lines 3 and 6 load patterns of vegetation indexes and samples points from text files. Line 7 retrieves the time series corresponding to the samples. Finally, line 8 calls the classifier on the samples using the patterns.

Figure 5.3 - Patterns (top) and samples (bottom) of NDVI time series for classification.



Listing 5.3 - Python code for classifying time series using Dynamic Time Warping.

```python
1 from dtw import *
2 from tools import *
3 patterns_ts = pd.read_json("examples/patterns.json", \
4                            orient='records')
5 patterns_ts["timeline"] = pd.to_datetime(patterns_ts["timeline"])
6 samples = pd.read_csv("examples/samples.csv")
7 samples_ts = wtss_get_time_series(samples)
8 classification = classifier_1nn(patterns_ts, samples_ts)
```

In summary, we joined data and analysis environments in order to plot, filter, and classify time series of Earth observation data by means of Jupyter notebooks and web services. This approach is flexible as users can use the same data and web services over different programming languages and analysis environments. For example, we setup another notebook using *R*, a statistical programming language. We do not describe this *R* notebook here, but the code is available on-line[2].

5.8 Conclusions

In this paper, we discussed how literate programming is being taking to the Web as interactive and collaborative analysis environments. We also showed how these environments are enhanced with web services and how both —environments and services— help scientists to prepare their analysis routines. We set up a Jupyter notebook in which we analyzed data retrieved by the Web Time Series Service. In this way, we showed how to display, filter, smooth and classify time series of vegetation indexes. This is a convenient for scientists not only to interact with time series of Earth observation data but also to prepare their analysis routines before running them on big Earth observation data platforms such as e-sensing.

Web services close the gap between big Earth observation data and analysis tools by means of collaborative environments for small amounts of data. As the amount of data to be processed increases, it is better to send the analysis routine to the data which is an ongoing effort at the e-sensing project.

Finally, we would like to remark that the aforementioned the Jupyter notebook, the Web Time Series Service, and the analysis routine are available on-line to everyone at `http://github.com/e-sensing/wgiss-py-webinar`.

[2] e-Sensing: Big Earth observation data analytics for Land Use and Land Cover change information `https://github.com/e-sensing/SITS_R_notebook`

6 CONCLUSION

In this thesis, we address scientific questions related to the Land Use and Land Cover component of the Earth system. Specifically, we focus on the use of Earth Observation data for information extraction, analysis, and distribution. Regarding data preparation, we investigate the performance of the available cloud detection algorithms on images of the Amazon forest. On the matter of transforming raw satellite data into deforestation maps, we propose a classification method based on time series and machine learning. Lastly, we test an approach for making exploratory data analysis data and methods available to scientists through the combination of literate programming with interpreted and high level programming languages running on web environments.

Regarding cloud detection algorithms, we found that Fmask performs better than other algorithms and so we used it later in our classifications. Cloud masking is of paramount importance downstream in the current trend of remote sensing which is moving to data cubes and analysis ready data. Cloud detection is a dynamic area. Since the publication of our paper, Fmask had two minor improvements, moving from version 4.0 to 4.2. Fmask also introduced the trend of supporting masking on images from different sensors and satellites. However, their algorithm is still largely coupled to the spectral bands offered by each satellite. On the other hand, the perspective of using time series and machine learning for cloud masking are brought by the MAJA and s2cloudless algorithms. We believe both approaches will eventually overcome the constraints imposed by each sensor and would be available to other Earth observation platforms, such as CBERS.

Next, we tested the use of remote sensing time series to detect clear-cut deforestation using machine learning networks and a linear mixture model and the raw bands of Landsat 8. These experiments probed the feasibility of our method and taught us two valuable lessons on how to refine it: We found the need of finer spatial and temporal resolution of deforestation monitoring systems. We also found that deep learning requires numerous samples which are costly to collect and validate. For that reason, our following classification was based on Sentinel-2 images with Random Forest on which our results resemble better visual classifications.

The use of time series and machine learning for analyzing massive datasets of Earth observation data is a promising approach for accurate and detailed understanding of the Earth system. In our experiments, we demonstrate that it is feasible to produce high-quality deforestation maps with a resolution of ten meters, despite the recurrent cloud cover over the Amazon forest. Our work brings valuable lessons for deforestation monitoring systems. For example, a monitoring system based on our experiments could use massive amounts of computing power to identify not only new but also recurrent deforestation, which would improve Carbon inventories that feed global models of planetary climate.

Our results also point to the need to make data, information, and methods available to scientist across operating systems, protocols, and languages. The amount of data involved, processing time, and analysis method complexity are barriers to the distribution of knowledge and perhaps more importantly, science reproducibility. In this thesis, we explore how collaborative analysis environments could encourage scientist to move from localized and small multitemporal datasets to large time series of Earth observation data online. As these technologies spread, putting together curated data collection along tested analysis algorithms, the missing element is scientific brain power for proposing and testing hypothesis.

Lastly, the future of our research lies in using our analysis on larger extents of the Amazon forest and also, in exploring beyond time series into using fast and scalable spatio-temporal analysis methods.

We believe the results in this theses are valuable for the currently on-going project Brazil Data Cube. This project is currently building arrays of Earth observation data similar to those we used in Chapter 2 and Chapter 4 for the five Brazilian biomes (*Amazônia*, *Cerrado*, *Mata Atlántica*, *Caatinga*, *Pampa*, and *Pantanal*). Our findings regarding cloud masking, classification methods, use of time series, and access to data and methods could improve their results and scientific reach, particularly for *Amazônia*.

REFERENCES

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; GOODFELLOW, I.; HARP, A.; IRVING, G.; ISARD, M.; JIA, Y.; JOZEFOWICZ, R.; KAISER, L.; KUDLUR, M.; LEVENBERG, J.; MANE, D.; MONGA, R.; MOORE, S.; MURRAY, D.; OLAH, C.; SCHUSTER, M.; SHLENS, J.; STEINER, B.; SUTSKEVER, I.; TALWAR, K.; TUCKER, P.; VANHOUCKE, V.; VASUDEVAN, V.; VIEGAS, F.; VINYALS, O.; WARDEN, P.; WATTENBERG, M.; WICKE, M.; YU, Y.; ZHENG, X. TensorFlow: large-scale machine learning on heterogeneous distributed systems. **arXiv:1603.04467 [cs]**, Mar 2016. 10

ACHARD, F.; STIBIG, H.-J.; EVA, H. D.; LINDQUIST, E. J.; BOUVET, A.; ARINO, O.; MAYAUX, P. Estimating tropical deforestation from Earth observation data. **Carbon Management**, v. 1, n. 2, p. 271–287, Dec 2010. ISSN 1758-3004. 38

ADARME, M.; FEITOSA, R. Q.; HAPP, P. N.; ALMEIDA, C. A. D.; GOMES, A. R. Evaluation of deep learning techniques for deforestation detection in the brazilian Amazon and Cerrado biomes from remote sensing imagery. **Remote Sensing**, v. 12, n. 6, p. 910, Jan 2020. 38, 46, 49, 50

ALMEIDA, C.; COUTINHO, A.; ESQUERDO, J.; ADAMI, M.; VENTURIERI, A.; DINIZ, C.; DESSAY, N.; DURIEUX, L.; GOMES, A. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. **Acta Amazonica**, v. 46, n. 3, p. 291–302, Sept 2016. 5, 16

ANONYMOUS. Reality check on reproducibility. **Nature**, v. 533, n. 7604, p. 437–437, May 2016. ISSN 0028-0836, 1476-4687. 56

______. Take action to stop the Amazon burning. **Nature**, v. 573, n. 7773, p. 163–163, Sept 2019. ISSN 0028-0836, 1476-4687. 37

ANTONELLI, A.; ZIZKA, A.; CARVALHO, F. A.; SCHARN, R.; BACON, C. D.; SILVESTRO, D.; CONDAMINE, F. L. Amazonia is the primary source of neotropical biodiversity. **Proceedings of the National Academy of Sciences**, v. 115, n. 23, p. 6034–6039, Jun 2018. ISSN 0027-8424, 1091-6490. 15

APPEL, M.; PEBESMA, E. On-demand processing of data cubes from satellite image collections with the gdalcubes library. **Data**, v. 4, n. 3, p. 1–16, 2019. 41

ARAGÃO, L. E. O. C.; ANDERSON, L. O.; FONSECA, M. G.; ROSAN, T. M.; VEDOVATO, L. B.; WAGNER, F. H.; SILVA, C. V. J.; SILVA JUNIOR, C. H. L.; ARAI, E.; AGUIAR, A. P.; BARLOW, J.; BERENGUER, E.; DEETER, M. N.; DOMINGUES, L. G.; GATTI, L.; GLOOR, M.; MALHI, Y.; MARENGO, J. A.; MILLER, J. B.; PHILLIPS, O. L.; SAATCHI, S. 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. **Nature Communications**, v. 9, n. 1, p. 536, Dec 2018. ISSN 2041-1723. 1

ARTAXO, P.; RIZZO, L. V.; PAIXÃO, M.; LUCCA, S. D.; OLIVEIRA, P. H.; LARA, L. L.; WIEDEMANN, K. T.; ANDREAE, M. O.; HOLBEN, B.; SCHAFER, J.; CORREIA, A. L.; PAULIQUEVIS, T. M. Aerosol particles in amazonia: their composition, role in the radiation balance, cloud formation, and nutrient cycles. In: KELLER, M.; BUSTAMANTE, M.; GASH, J.; DIAS, P. S. (Ed.). **Amazonia and global change**. [S.l.]: American Geophysical Union (AGU), 2009. p. 233–250. ISBN 978-1-118-67034-7. 17

ASNER, G. P. Cloud cover in Landsat observations of the brazilian Amazon. **International Journal of Remote Sensing**, v. 22, n. 18, p. 3855–3862, Jan 2001. ISSN 0143-1161, 1366-5901. 17, 18

______. Automated mapping of tropical deforestation and forest degradation: CLASlite. **Journal of Applied Remote Sensing**, v. 3, n. 1, p. 033543, Aug 2009. ISSN 1931-3195. 46

ASNER, G. P.; KNAPP, D. E.; BROADBENT, E. N.; OLIVEIRA, P. J. C.; KELLER, M.; SILVA, J. N. Selective logging in the brazilian Amazon. **Science**, v. 310, n. 5747, p. 480–482, 2005. 2

ASSIS, L. F.; RIBEIRO, G.; FERREIRA, K. R.; VINHAS, L.; LLAPA, E.; SANCHEZ, A.; MAUS, V.; CAMARA, G. Big data streaming for remote sensing time series analytics using MapReduce. In: **BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 17., 2016. Proceedings...** [S.l.: s.n.], 2016. 57

ASSUNÇÃO, J.; GANDOUR, C.; ROCHA, R. Deforestation slowdown in the Brazilian Amazon: prices or policies? **Environment and Development Economics**, v. 20, n. 6, p. 697–722, 2015. 16

ATKINSON, P. M.; JEGANATHAN, C.; DASH, J.; ATZBERGER, C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. **Remote Sensing of Environment**, v. 123, p. 400–417, 2012. 59

ATZBERGER, C.; EILERS, P. H. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. **International Journal of Remote Sensing**, v. 32, n. 13, p. 3689–3709, 2011. 60

AZEVEDO-RAMOS, C.; MOUTINHO, P.; ARRUDA, V. L.; STABILE, M. C.; ALENCAR, A.; CASTRO, I.; RIBEIRO, J. P. Lawless land in no man's land: The undesignated public forests in the Brazilian Amazon. **Land Use Policy**, Elsevier, v. 99, n. June, p. 104863, 2020. ISSN 02648377. Available from: <https://doi.org/10.1016/j.landusepol.2020.104863>. 1

BAETENS, L.; DESJARDINS, C.; HAGOLLE, O. Validation of Copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. **Remote Sensing**, v. 11, n. 4, p. 433, Jan 2019. 17, 22, 30, 31, 33, 34

BAKER, M. 1,500 scientists lift the lid on reproducibility. **Nature**, v. 533, n. 7604, p. 452–454, May 2016. ISSN 0028-0836, 1476-4687. 56

BARBER, C. P.; COCHRANE, M. A.; SOUZA, C. M.; LAURANCE, W. F. Roads, deforestation, and the mitigating effect of protected areas in the Amazon. **Biological Conservation**, v. 177, p. 203–209, Sept 2014. ISSN 00063207. 1

BARLOW, J.; BERENGUER, E.; CARMENTA, R.; FRANÇA, F. Clarifying amazonia's burning crisis. **Global Change Biology**, v. 2019, p. gcb.14872, Nov 2019. ISSN 1354-1013. 1

BARLOW, J.; PERES, C. A. Fire-mediated dieback and compositional cascade in an Amazonian forest. **Philosphical Transactions Royal Society London B Biological Sciences**, v. 363, p. 1787–1794, 2008. 1

BEGOTTI, R. A.; PERES, C. A. Rapidly escalating threats to the biodiversity and ethnocultural capital of Brazilian Indigenous Lands. **Land Use Policy**, v. 96, p. 104694, Jul 2020. ISSN 02648377. 37

BELGIU, M.; DRAGUT, L. Random forest in remote sensing: a review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016. 42

BELL, G.; HEY, T.; SZALAY, A. Computer science: Beyond the Data Deluge. **Science**, v. 323, n. 5919, p. 1297–1298, Mar 2009. ISSN 0036-8075, 1095-9203. 56

BENGIO, Y. Practical recommendations for gradient-based training of deep architectures. **arXiv:1206.5533** [cs], Sept 2012. 9

BERNDT, D.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: FAYYAD, U. M.; UTHURUSAMY, R. (Ed.). **Workshop on knowledge knowledge discovery in databases**. AAAI Press, 1994. v. 398, p. 359–370. ISBN 0-929280-73-3. Available from: <http://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf>. 60

BOYD, D.; CRAWFORD, K. Critical questions for big data. **Information, Communication & Society**, v. 15, n. 5, p. 662–679, Jun 2012. ISSN 1369-118X. 56

BRANCALION, P. H. S.; DE ALMEIDA, D. R. A.; VIDAL, E.; MOLIN, P. G.; SONTAG, V. E.; SOUZA, S. E. X. F.; SCHULZE, M. D. Fake legal logging in the brazilian Amazon. **Science Advances**, v. 4, n. 8, p. eaat1192, Aug 2018. ISSN 2375-2548. 1

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125. 42

BROWN, J. C.; KASTENS, J. H.; COUTINHO, A. C.; VICTORIA, D. d. C.; BISHOP, C. R. Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. **Remote Sensing of Environment**, v. 130, p. 39–50, 2013. 16

CAMARA, G.; ASSIS, L. F.; RIBEIRO, G.; FERREIRA, K. R.; LLAPA, E.; VINHAS, L. Big earth observation data analytics: matching requirements to system architectures. In: **ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON ANALYTICS FOR BIG GEOSPATIAL DATA, 5, 2016. Proceedings...** Burlingname, CA, USA: ACM, 2016. p. 1–6. 38, 57

CAMARA, G.; MACIEL, A.; MAUS, V.; VINHAS, L.; SANCHEZ, A. Using dynamic geospatial ontologies to support information extraction from big Earth Events as key concepts for describing land use change. In: **GISCIENCE, 2016. Proceedings...** Montreal, Canada: [s.n.], 2016. 57

CAMARA, G.; SIMOES, R.; ANDRADE, P. R.; MAUS, V.; SÁNCHEZ, A.; ASSIS, L. F. F. G. D.; LORENALVES; YWATACARVALHO; MACIEL, A. M.; VINHAS, L.; QUEIROZ, G. R. D. **E-Sensing/Sits: Version 1.12.5**. Dec 2018. Zenodo. 42

CANTON, T. D.; NITZ, A. H.; LUNDGREN, A. P.; NIELSEN, A. B.; BROWN, D. A.; DENT, T.; HARRY, I. W.; KRISHNAN, B.; MILLER, A. J.; WETTE, K.; WIESNER, K.; WILLIS, J. L. Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors. **Physical Review D**, v. 90, n. 8, p. 082004, Oct 2014. ISSN 1550-7998, 1550-2368. 59

CECCHINI, M. A.; MACHADO, L. A. T.; ANDREAE, M. O.; MARTIN, S. T.; ALBRECHT, R. I.; ARTAXO, P.; BARBOSA, H. M. J.; BORRMANN, S.; FÜTTERER, D.; JURKAT, T.; MAHNKE, C.; MINIKIN, A.; MOLLEKER, S.; PÖHLKER, M. L.; PÖSCHL, U.; ROSENFELD, D.; VOIGT, C.; WEINZIERL, B.; WENDISCH, M. Sensitivities of amazonian clouds to aerosols and updraft speed. **Atmospheric Chemistry and Physics**, v. 17, n. 16, p. 10037–10050, Aug 2017. ISSN 1680-7324. 17

CEOS. **The CEOS data dube. three-year work plan 2016-2018**. Committee on Earth Observation Satellites, 2016. Available from: <http://ceos.org/document_management/Ad_Hoc_Teams/SDCG_for_GFOI/Meetings/SDCG-10/Cube3-YearWorkPlan-v1.0.pdf>. 57

CHAZDON, R. L. et al. When is a forest a forest? forest concepts and definitions in the era of forest and landscape restoration. **Ambio**, v. 45, n. 5, p. 538–50, 2016. 1

CHINCHOR, N. MUC-4 Evaluation metrics. In: **MESSAGE UNDERSTANDING CONFERENCE, 4., 1992. Proceedings...** McLean, Virginia: [s.n.], 1992. 26

CHOLLET, F. et al. **Keras (2015)**. 2017. Available from: <https://keras.io>. 10

CIBULA, W. G.; ZETKA, E. F.; RICKMAN, D. L. Response of thematic mapper bands to plant water stress. **International Journal of Remote Sensing**, v. 13, n. 10, p. 1869–1880, Jul 1992. ISSN 0143-1161, 1366-5901. 41

CLAVERIE, M.; JU, J.; MASEK, J. G.; DUNGAN, J. L.; VERMOTE, E. F.; ROGER, J.-C.; SKAKUN, S. V.; JUSTICE, C. The harmonized Landsat and Sentinel-2 surface reflectance data set. **Remote Sensing of Environment**, v. 219, p. 145–161, Dec 2018. ISSN 00344257. 13, 16

CLEVERS, J. G. P. W.; KOOISTRA, L. Using hyperspectral remote sensing data for retrieving canopy chlorophyll and nitrogen content. **IEEE Journal of**

Selected Topics in Applied Earth Observations and Remote Sensing, v. 5, n. 2, p. 574–583, Apr 2012. ISSN 1939-1404, 2151-1535. 46

DAVIDSON, E. A.; ARAUJO, A. C.; ARTAXO, P.; BALCH, J. K.; BROWN, I. F.; BUSTAMANTE, M. M. C.; COE, M. T.; DEFRIES, R. S.; KELLER, M.; LONGO, M.; MUNGER, J. W.; SCHROEDER, W.; SOARES-FILHO, B.; SOUZA, C. M.; WOFSY, S. C. The Amazon basin in transition. **Nature**, v. 481, p. 321–328, 2012. 18

DEFOURNY, P.; BONTEMPS, S.; BELLEMANS, N.; CARA, C.; DEDIEU, G.; GUZZONATO, E.; HAGOLLE, O.; INGLADA, J.; NICOLA, L.; RABAUTE, T.; SAVINAUD, M.; UDROIU, C.; VALERO, S.; BÉGUÉ, A.; DEJOUX, J.-F.; HARTI, A. E.; EZZAHAR, J.; KUSSUL, N.; LABBASSI, K.; LEBOURGEOIS, V.; MIAO, Z.; NEWBY, T.; NYAMUGAMA, A.; SALH, N.; SHELESTOV, A.; SIMONNEAUX, V.; TRAORE, P. S.; TRAORE, S. S.; KOETZ, B. Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the Sen2-Agri automated system in various cropping systems around the world. **Remote Sensing of Environment**, v. 221, p. 551–568, Feb 2019. ISSN 00344257. 21

DIAS, L. C. P.; PIMENTA, F. M.; SANTOS, A. B.; COSTA, M. H.; LADLE, R. J. Patterns of land use, extensification, and intensification of Brazilian agriculture. **Global Change Biology**, v. 22, n. 8, p. 2887–2903, Aug 2016. ISSN 13541013. 2

DINIZ, C. G.; SOUZA, A. A. d. A.; SANTOS, D. C.; DIAS, M. C.; LUZ, N. C. da; MORAES, D. R. V. de; MAIA, J. S. A.; GOMES, A. R.; NARVAES, I. d. S.; VALERIANO, D. M.; MAURANO, L. E. P.; ADAMI, M. DETER-B: The new Amazon near real-time deforestation detection system. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 8, n. 7, p. 3619–3628, Jul 2015. ISSN 1939-1404, 2151-1535. 1, 38

DRUSCH, M.; BELLO, U. D.; CARLIER, S.; COLIN, O.; FERNANDEZ, V.; GASCON, F.; HOERSCH, B.; ISOLA, C.; LABERINTI, P.; MARTIMORT, P.; MEYGRET, A.; SPOTO, F.; SY, O.; MARCHESE, F.; BARGELLINI, P. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. **Remote Sensing of Environment**, v. 120, p. 25–36, May 2012. ISSN 0034-4257. 16, 18

DURIEUX, L. The impact of deforestation on cloud cover over the Amazon arc of deforestation. **Remote Sensing of Environment**, v. 86, n. 1, p. 132–140, Jun 2003. ISSN 00344257. 17

EILERS, P. H. C. A perfect smoother. **Analytical Chemistry**, v. 75, n. 14, p. 3631–3636, Jul 2003. ISSN 0003-2700, 1520-6882. 60

ESCOBAR, H. Brazilian president attacks deforestation data. **Science**, v. 365, n. 6452, p. 419–419, Aug 2019. ISSN 0036-8075, 1095-9203. 5

EXBRAYAT, J.-F.; LIU, Y. Y.; WILLIAMS, M. Impact of deforestation and climate on the Amazon Basin's above-ground biomass during 1993–2012. **Scientific Reports**, v. 7, n. 1, p. 15615, Dec 2017. ISSN 2045-2322. 37

FEARNSIDE, P. M. Environmental and social impacts of hydroelectric dams in brazilian Amazonia: implications for the aluminum industry. **World Development**, v. 77, p. 48–65, Jan 2016. ISSN 0305750X. 1

FERRANTE, L.; FEARNSIDE, P. M. The Amazon: biofuels plan will drive deforestation. **Nature**, v. 577, n. 7789, p. 170–170, Jan 2020. ISSN 0028-0836, 1476-4687. 37

FINER, M.; NOVOA, S.; WEISSE, M. J.; PETERSEN, R.; MASCARO, J.; SOUTO, T.; STEARNS, F.; MARTINEZ, R. G. Combating deforestation: from satellite to intervention. **Science**, v. 360, n. 6395, p. 1303–1305, Jun 2018. ISSN 0036-8075, 1095-9203. 38

FOGA, S.; SCARAMUZZA, P. L.; GUO, S.; ZHU, Z.; DILLEY, R. D.; BECKMANN, T.; SCHMIDT, G. L.; DWYER, J. L.; HUGHES, M. J.; LAUE, B. Cloud detection algorithm comparison and validation for operational Landsat data products. **Remote Sensing of Environment**, v. 194, p. 379–390, Jun 2017. ISSN 00344257. 21, 23

FOLEY, J. A.; DEFRIES, R.; ASNER, G. P.; BARFORD, C.; BONAN, G.; CARPENTER, S. R.; CHAPIN, F. S.; COE, M. T.; DAILY, G. C.; GIBBS, H. K.; HELKOWSKI, J. H.; HOLLOWAY, T.; HOWARD, E. A.; KUCHARIK, C. J.; MONFREDA, C.; PATZ, J. A.; PRENTICE, I. C.; RAMANKUTTY, N.; SNYDER, P. K. Global consequences of land use. **Science**, v. 309, n. 5734, p. 570–574, 2005. 1

FRANTZ, D.; HASS, E.; UHL, A.; STOFFELS, J.; HILL, J. Improvement of the Fmask algorithm for Sentinel-2 images: separating clouds from bright surfaces based on parallax effects. **Remote Sensing of Environment**, v. 215, p. 471–481, Sept 2018. ISSN 00344257. 17, 21

GASCON, F.; BOUZINAC, C.; THEPAUT, O.; JUNG, M.; FRANCESCONI, B.; LOUIS, J.; LONJOU, V.; LAFRANCE, B.; MASSERA, S.; GAUDEL-VACARESSE, A.; LANGUILLE, F.; ALHAMMOUD, B.; VIALLEFONT, F.; PFLUG, B.; BIENIARZ, J.; CLERC, S.; PESSIOT, L.; TREMAS, T.; CADAU, E.; BONIS, R. D.; ISOLA, C.; MARTIMORT, P.; FERNANDEZ, V. Copernicus Sentinel-2A calibration and products validation status. **Remote Sensing**, v. 9, n. 6, 2017. ISSN 20724292. 19, 21, 22, 41

GIBBS; RUESCH, A. S.; ACHARD, F.; CLAYTON, M. K.; HOLMGREN, P.; RAMANKUTTY, N.; FOLEY, J. A. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. **Proceedings of the National Academy of Sciences**, v. 107, n. 38, p. 16732–16737, 2010. 15

GIBBS, H.; MUNGER, J.; L'ROE, J.; BARRETO, P.; PEREIRA, R.; CHRISTIE, M.; AMARAL, T.; WALKER, N. Did ranchers and slaughterhouses respond to zero-deforestation agreements in the Brazilian Amazon? **Conservation Letters**, v. 9, p. 32–42, 2015. 16

GIBBS, H. K.; RAUSCH, L.; MUNGER, J.; SCHELLY, I.; MORTON, D. C.; NOOJIPADY, P.; SOARES-FILHO, B.; BARRETO, P.; MICOL, L.; WALKER, N. F. Brazil's soy moratorium. **Science**, v. 347, n. 6220, p. 377–378, Jan 2015. ISSN 0036-8075, 1095-9203. 47

GITELSON, A.; MERZLYAK, M. N. Quantitative estimation of chlorophyll-a using reflectance spectra: experiments with autumn chestnut and maple leaves. **Journal of Photochemistry and Photobiology B: Biology**, v. 22, n. 3, p. 247–252, Mar 1994. ISSN 10111344. 46

GIULIANI, G.; CAMARA, G.; KILLOUGH, B.; MINCHIN, S. Earth observation open science: enhancing reproducible science using data cubes. **Data**, v. 4, n. 4, p. 147, Dec 2019. 38

GOLLNOW, F.; HISSA, L. d. B. V.; RUFIN, P.; LAKES, T. Property-level direct and indirect deforestation for soybean production in the Amazon region of Mato Grosso, Brazil. **Land Use Policy**, v. 78, p. 377–385, 2018. ISSN 0264-8377. 1

GOMEZ, C.; WHITE, J. C.; WULDER, M. A. Optical remotely sensed time series data for land cover classification: a review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 116, p. 55–72, 2016. ISSN 0924-2716. 38

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT Press, 2016. 9

GREWAL, M. S.; ANDREWS, A. P. Applications of Kalman Filtering in Aerospace 1960 to the Present [Historical Perspectives]. **IEEE Control Systems**, v. 30, n. 3, p. 69–78, Jun 2010. ISSN 1066-033X, 1941-000X. 60

GRIFFITHS, P.; LINDEN, S.; KUEMMERLE, T.; HOSTERT, P. A pixel-based Landsat compositing algorithm for large area land cover mapping. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 6, n. 5, p. 2088–2101, 2013. 16

GRINAND, C.; RAKOTOMALALA, F.; GOND, V.; VAUDRY, R.; BERNOUX, M.; VIEILLEDENT, G. Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier. **Remote Sensing of Environment**, v. 139, p. 68–80, Dec 2013. ISSN 00344257. 38, 46, 49, 50

HAGOLLE, O.; HUC, M.; AUER, S.; RICHTER, R.; RICHTER, R. **MAJA algorithm theoretical basis document**. [S.l.]: CNES, 2017. 17, 21, 22

HAGOLLE, O.; HUC, M.; PASCUAL, D. V.; DEDIEU, G. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and Sentinel-2 images. **Remote Sensing**, v. 7, n. 3, p. 2668–2691, Mar 2015. 22

HAMUNYELA, E.; VERBESSELT, J.; HEROLD, M. Using spatial context to improve early detection of deforestation from Landsat time series. **Remote Sensing of Environment**, v. 172, p. 126–138, Jan 2016. ISSN 00344257. 46, 49, 50

HANSEN, M. C.; LOVELAND, T. R. A review of large area monitoring of land cover change using Landsat data. **Remote Sensing of Environment**, v. 122, p. 66–74, 2012. ISSN 0034-4257. 1

HANSEN, M. C.; POTAPOV, P. V.; MOORE, R.; HANCHER, M.; TURUBANOVA, S. A.; TYUKAVINA, A.; THAU, D.; STEHMAN, S. V.; GOETZ, S. J.; LOVELAND, T. R.; KOMMAREDDY, A.; EGOROV, A.; CHINI, L.; JUSTICE, C. O.; TOWNSHEND, J. R. G. High-resolution global maps of 21st-century forest cover change. **Science**, v. 342, n. 6160, p. 850–853, 2013. 16, 38, 47

HANSEN, M. C.; WANG, L.; SONG, X.-P.; TYUKAVINA, A.; TURUBANOVA, S.; POTAPOV, P. V.; STEHMAN, S. V. The fate of tropical forest fragments. **Science Advances**, v. 6, n. 11, p. eaax8574, Mar 2020. ISSN 2375-2548. 1, 38, 51

HASTIE, T.; TIBSHIRANI, R.; J, F. **The elements of statistical learning. data mining, inference, and prediction**. New York: Springer, 2009. 34

HEIDEMAN, M.; JOHNSON, D.; BURRUS, C. Gauss and the history of the fast fourier transform. **IEEE ASSP Magazine**, v. 1, n. 4, p. 14–21, Oct 1984. ISSN 0740-7467. 60

HEY, T.; TANSLEY, S.; TOLLE, K. M. Jim gray on eScience: a transformed scientific method. 2009. Available from: <http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf>. 3, 56

HUETE, A.; JACKSON, R.; POST, D. Spectral response of a plant canopy with different soil backgrounds. **Remote Sensing of Environment**, v. 17, n. 1, p. 37–53, Feb 1985. ISSN 00344257. 41, 59

IHAKA, R.; GENTLEMAN, R. R: A language for data analysis and graphics. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 299, Sept 1996. ISSN 10618600. 42, 59

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Municipal livestock research**. Rio de Janeiro: IBGE, 2018. 39

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE. **PRODES - incremento anual de area desmatada no Cerrado brasileiro**. São José dos Campos: INPE, 2019. 15

______. ______. São José dos Campos: INPE, 2020. 37, 39, 47

ISRAEL, G. D. **Determining sample size**. University of Florida, 1992. 1–5 p. 23

JAKIMOW, B.; GRIFFITHS, P.; LINDEN, S.; HOSTERT, P. Mapping pasture management in the Brazilian Amazon from dense Landsat time series. **Remote Sensing of Environment**, v. 205, p. 453–468, Feb 2018. ISSN 00344257. 16

JAKUBAUSKAS., M. E.; LEGATES., D. R.; KASTENS., J. H. Harmonic analysis of time-series AVHRR NDVI data. **Photogrammetric Engineering and Remote Sensing**, v. 67, n. 4, p. 461–470., 2001. 60

JENSEN, J. R. **Remote sensing of the environment: an Earth resource perspective**. Delhi, India: Pearson Education, 2009. ISBN 978-81-317-1680-9. 46

JIANG, Z.; HUETE, A. R.; DIDAN, K.; MIURA, T. Development of a two-band enhanced vegetation index without a blue band. **Remote Sensing of Environment**, v. 112, n. 10, p. 3833–3845, Oct 2008. ISSN 0034-4257. 59

KALAMANDEEN, M.; GLOOR, E.; MITCHARD, E.; QUINCEY, D.; ZIV, G.; SPRACKLEN, D.; SPRACKLEN, B.; ADAMI, M.; ARAGÃO, L. E. O. C.; GALBRAITH, D. Pervasive rise of small-scale deforestation in Amazonia. **Scientific Reports**, v. 8, n. 1, p. 1600, Dec 2018. ISSN 2045-2322. 1, 38, 51, 52

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: a highly efficient gradient boosting decision tree. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in neural information processing systems**. [S.l.: s.n.], 2017. p. 3146–3154. 22

KLEYNHANS, W.; OLIVIER, J. C.; WESSELS, K. J.; SALMON, B. P.; van den BERGH, F.; STEENKAMP, K. Detecting land cover change using an extended Kalman filter on MODIS NDVI time-series data. **IEEE Geoscience and Remote Sensing Letters**, v. 8, n. 3, p. 507–511, May 2011. ISSN 1545-598X, 1558-0571. 60

KLUYVER, T.; RAGAN-KELLEY, B.; PÉREZ, F.; GRANGER, B.; BUSSONNIER, M.; FREDERIC, J.; KELLEY, K.; HAMRICK, J.; GROUT, J.; CORLAY, S.; IVANOV, P.; AVILA, D.; ABDALLA, S.; WILLING, C. Jupyter notebooks—a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (Ed.). **Positioning and Power in Academic Publishing: Players, Agents and Agendas**. [S.l.]: IOS, 2016. p. 87–90. ISBN 9781614996491. 59

KNUTH, D. E. Literate programming. **The Computer Journal**, v. 27, n. 2, p. 97–111, Feb 1984. ISSN 0010-4620, 1460-2067. 3, 56, 58

LEVER, J.; KRZYWINSKI, M.; ALTMAN, N. Model selection and overfitting. **Nature Methods**, v. 13, n. 9, p. 703–704, Sept 2016. ISSN 1548-7091, 1548-7105. 46

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A.; CHENG,

T. Geospatial big data handling theory and methods: a review and research challenges. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 115, p. 119–133, May 2016. ISSN 0924-2716. 56

LOUIS, J.; DEBAECKER, V.; PFLUG, B.; MAIN-KNORN, M.; BIENIARZ, J.; MUELLER-WILM, U.; CADAU, E.; GASCON, F. SENTINEL-2 Sen2Cor: L2A processor for users. In: **LIVING PLANET SYMPOSIUM, MESSAGE UNDERSTANDING CONFERENCE, 2016. Proceedings...** [S.l.]: ESA, 2016. p. 8. 17, 21

LU, M.; PEBESMA, E.; SANCHEZ, A.; VERBESSELT, J. Spatio-temporal change detection from multidimensional arrays: detecting deforestation from MODIS time series. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 117, p. 227–236, Jul 2016. ISSN 09242716. 57

MACIEL, A.; CAMARA, G.; VINHAS, L.; PICOLI, M.; BEGOTTI, R.; ASSIS, L. F. A spatiotemporal calculus for reasoning about land-use trajectories. **International Journal of Geographical Information Science**, v. 33, n. 1, p. 176–192, 2019. 57

MAHECHA, M. D.; GANS, F.; BRANDT, G.; CHRISTIANSEN, R.; CORNELL, S. E.; FOMFERRA, N.; KRAEMER, G.; PETERS, J.; BODESHEIM, P.; CAMPS-VALLS, G.; DONGES, J. F.; DORIGO, W.; ESTUPINAN-SUAREZ, L. M.; GUTIERREZ-VELEZ, V. H.; GUTWIN, M.; JUNG, M.; LONDOÑO, M. C.; MIRALLES, D. G.; PAPASTEFANOU, P.; REICHSTEIN, M. Earth system data cubes unravel global multivariate dynamics. **Earth System Dynamics**, v. 11, n. 1, p. 201–234, Feb 2020. ISSN 2190-4979. 38

MAURANO, L. E. P.; ESCADA, M. I. S. Comparação dos dados produzidos pelo PRODES versus dados do mapbiomas para o bioma Amazônia. **SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 19., 2019. Anais...**, São José dos Campos: INPE, 2019, p. 735–738, 2019. 12

MAUS, V.; CAMARA, G.; CARTAXO, R.; SANCHEZ, A.; RAMOS, F. M.; QUEIROZ, G. R. A Time-weighted dynamic time warping method for land-use and land-cover mapping. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 9, n. 8, p. 3729–3739, 2016. 57

MILODOWSKI, D. T.; MITCHARD, E. T. A.; WILLIAMS, M. Forest loss maps from regional satellite monitoring systematically underestimate deforestation in two rapidly changing parts of the Amazon. **Environmental Research Letters**, v. 12, n. 9, p. 094003, Sept 2017. ISSN 1748-9326. 38

MONTIBELLER, B.; KMOCH, A.; VIRRO, H.; MANDER, Ü.; UUEMAA, E. Increasing fragmentation of forest cover in Brazil's Legal Amazon from 2001 to 2017. **Scientific Reports**, v. 10, n. 1, p. 5803, Dec 2020. ISSN 2045-2322. 1

MUELLER, H.; GRIFFITHS, P.; HOSTERT, P. Long-term deforestation dynamics in the Brazilian Amazon-Uncovering historic frontier development along the Cuiaba-Santarem highway. **International Journal of Applied Earth Observation and Geoinformation**, v. 44, p. 61–69, 2016. 1, 39

MUELLER-WILM, U. **Sen2Cor 2.8 software release note**. Frascati, 2019. 22

NEPSTAD, D.; MCGRATH, D.; STICKLER, C.; ALENCAR, A.; AZEVEDO, A.; SWETTE, B.; BEZERRA, T.; DIGIANO, M.; SHIMADA, J.; SEROA DA MOTTA, R.; ARMIJO, E.; CASTELLO, L.; BRANDO, P.; HANSEN, M. C.; MCGRATH-HORN, M.; CARVALHO, O.; HESS, L. Slowing Amazon deforestation through public policy and interventions in beef and soy supply chains. **Science**, v. 344, n. 6188, p. 1118–1123, 2014. 15

OLOFSSON, P.; FOODY, G. M.; HEROLD, M.; STEHMAN, S. V.; WOODCOCK, C. E.; WULDER, M. A. Good practices for estimating area and assessing accuracy of land change. **Remote Sensing of Environment**, v. 148, p. 42–57, 2014. 42, 44, 48

OMETTO, J.; AGUIAR, A.; ASSIS, T.; SOLER, L.; VALLE, P.; TEJADA, G.; LAPOLA, D.; MEIR, P. Amazon forest biomass density maps: tackling the uncertainty in carbon emission estimates. **Climatic Change**, v. 124, n. 3, p. 545–560, 2014. 15

PARENTE, L.; FERREIRA, L.; FARIA, A.; NOGUEIRA, S.; ARAUJO, F.; TEIXEIRA, L.; HAGEN, S. Monitoring the Brazilian pasturelands: a new mapping approach based on the Landsat 8 spectral and temporal domains. **International Journal of Applied Earth Observation and Geoinformation**, v. 62, p. 135–143, 2017. 47

PEKEL, J. F.; COTTAM, A.; GORELICK, N.; BELWARD, A. S. High-resolution mapping of global surface water and its long-term changes. **Nature**, v. 540, p. 418–422, 2016. 21

PEREZ, F.; GRANGER, B. E. IPython: a system for interactive scientific computing. **Computing in Science & Engineering**, v. 9, n. 3, p. 21–29, 2007. ISSN 1521-9615. 3, 56

PICOLI, M.; CAMARA, G.; SANCHES, I.; SIMOES, R.; CARVALHO, A.; MACIEL, A.; COUTINHO, A.; ESQUERDO JULIO A ND ANTUNES, J.; BEGOTTI, R. A.; ARVOR, D.; ALMEIDA, C. Big earth observation time series analysis for monitoring Brazilian agriculture. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 145, p. 328–339, 2018. 16, 38

PICOLI, M. C. A.; RORATO, A.; LEITÃO, P.; CAMARA, G.; MACIEL, A.; HOSTERT, P.; SANCHES, I. D. Impacts of public and private sector policies on soybean and pasture expansion in Mato Grosso—Brazil from 2001 to 2017. **Land**, v. 9, n. 1, p. 20, Jan 2020. 37, 47

POSCHL, U.; MARTIN, S. T.; SINHA, B.; CHEN, Q.; GUNTHE, S. S.; HUFFMAN, J. A.; BORRMANN, S.; FARMER, D. K.; GARLAND, R. M.; HELAS, G.; JIMENEZ, J. L.; KING, S. M.; MANZI, A.; MIKHAILOV, E.; PAULIQUEVIS, T.; PETTERS, M. D.; PRENNI, A. J.; ROLDIN, P.; ROSE, D.; SCHNEIDER, J.; SU, H.; ZORN, S. R.; ARTAXO, P.; ANDREAE, M. O. Rainforest aerosols as biogenic nuclei of clouds and precipitation in the Amazon. **Science**, v. 329, n. 5998, p. 1513–1516, Sept 2010. ISSN 0036-8075, 1095-9203. 17

QIN, Y.; XIAO, X.; DONG, J.; ZHANG, Y.; WU, X.; SHIMABUKURO, Y.; ARAI, E.; BIRADAR, C.; WANG, J.; ZOU, Z.; LIU, F.; SHI, Z.; DOUGHTY, R.; MOORE, B. Improved estimates of forest cover and loss in the Brazilian Amazon in 2000–2017. **Nature Sustainability**, v. 2, n. 8, p. 764–772, 2019. 1

QIU, S.; LIN, Y.; SHANG, R.; ZHANG, J.; MA, L.; ZHU, Z. Making Landsat time series consistent: evaluating and improving Landsat analysis ready data. **Remote Sensing**, v. 11, n. 1, p. 51, 2019. ISSN 2072-4292. 17, 21, 27, 31, 32, 34

QIU, S.; ZHU, Z.; HE, B. **Fmask 4.0 handbook**. May 2018. 23

REDMONK programming language rankings: June 2015. Available from: <https://redmonk.com/sogrady/2015/07/01/language-rankings-6-15>. 59

RICHARDS, P. What drives indirect land use change? how Brazil's agriculture sector influences frontier deforestation. **Annals of the Association of American Geographers Association of American Geographers**, v. 105, n. 5, p. 1026–1040, 2015. 37

RICHARDS, P.; ARIMA, E.; VANWEY, L.; COHN, A.; BHATTARAI, N. Are Brazil's deforesters avoiding detection? **Conservation Letters**, v. 10, n. 4, p. 470–476, Jul 2017. ISSN 1755-263X, 1755-263X. 1, 38, 51

ROBERTS, G. C.; ANDREAE, M. O.; ZHOU, J.; ARTAXO, P. Cloud condensation nuclei in the Amazon Basin: "marine" conditions over a continent? **Geophysical Research Letters**, v. 28, n. 14, p. 2807–2810, 2001. ISSN 1944-8007. 17

RUFIN, P.; MUELLER, H.; PFLUGMACHER, D.; HOSTERT, P. Land use intensity trajectories on Amazonian pastures derived from Landsat time series. **International Journal of Applied Earth Observation and Geoinformation**, v. 41, p. 1–10, 2015. 16

SANCHEZ, A. **Source code for: deforestation Sentinel-2**. Zenodo, 2020. Available from: <https://doi.org/10.5281/zenodo.3932012>. 53

SANCHEZ, A.; PICOLI, M.; SIMOES, R.; CAMARA, G.; ANDRADE, P.; FERREIRA, K. data set, **Deforestation maps using time series of Sentinel-2A images in Amazonia, between Brazil and Bolivia, in 2019**. PANGAEA, 2020. Available from: <https://doi.pangaea.de/10.1594/PANGAEA.921387>. 53

SANCHEZ, A. H.; PICOLI, M. C. A.; CAMARA, G.; ANDRADE, P. R.; CHAVES, M. E. D.; LECHLER, S.; SOARES, A. R.; MARUJO, R. F. B.; SIMÕES, R. E. O.; FERREIRA, K. R.; QUEIROZ, G. R. Comparison of cloud cover detection algorithms on Sentinel–2 images of the Amazon tropical forest. **Remote Sensing**, v. 12, n. 8, p. 1284, Jan 2020. 40

SCHULTZ, M.; CLEVERS, J. G.; CARTER, S.; VERBESSELT, J.; AVITABILE, V.; QUANG, H. V.; HEROLD, M. Performance of vegetation indices from Landsat time series in deforestation monitoring. **International Journal of Applied Earth Observation and Geoinformation**, v. 52, p. 318–327, Oct 2016. ISSN 03032434. 41, 46, 49, 50

SEGAL-ROZENHAIMER, M.; LI, A.; DAS, K.; CHIRAYATH, V. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). **Remote Sensing of Environment**, v. 237, p. 111446, Feb 2020. ISSN 00344257. 32

SEYMOUR, F.; HARRIS, N. L. Reducing tropical deforestation. **Science**, v. 365, n. 6455, p. 756–757, Aug 2019. ISSN 0036-8075, 1095-9203. 5

SHIMABUKURO, Y. E.; SANTOS, J. R.; FORMAGGIO, A. R.; DUARTE, V.; RUDORFF, B. F. T. The Brazilian Amazon monitoring program: PRODES and DETER projects. In: ACHARD, F.; HANSEN, M. C. (Ed.). **Global forest monitoring from earth observation**. [S.l.]: CRC Press, 2012. p. 354. 2, 5, 16, 38

SIMOES, R.; CAMARA, G.; ANDRADE, P.; CARVALHO, A. Y.; SANTOS, L.; FERREIRA, K.; MAUS, V.; QUEIROZ, G. **SITS: data analysis and machine learning using satellite image time series**. São José dos Campos; [s.n.], 2019. 10

SIMOES, R.; PICOLI, M. C. A.; CAMARA, G.; MACIEL, A.; SANTOS, L.; ANDRADE, P. R.; SÁNCHEZ, A.; FERREIRA, K.; CARVALHO, A. Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017. **Scientific Data**, v. 7, n. 1, p. 34, Dec 2020. ISSN 2052-4463. 38

SOTERRONI, A. C.; MOSNIER, A.; CARVALHO, A. X. Y.; CAMARA, G.; OBERSTEINER, M.; ANDRADE, P. R.; SOUZA, R. C.; BROCK, R.; PIRKER, J.; KRAXNER, F.; HAVLIK, P.; KAPOS, V.; ERMGASSEN, E.; VALIN, H.; RAMOS, F. M. Future environmental and agricultural impacts of Brazil's Forest Code. **Environmental Research Letters**, v. 13, n. 7, p. 074021, Jul 2018. ISSN 1748-9326. 15

SOUSA, D.; SMALL, C. Global cross-calibration of Landsat spectral mixture models. **Remote Sensing of Environment**, v. 192, p. 139–149, Apr 2017. ISSN 00344257. 8

SOUZA JUNIOR, C.; SIQUEIRA, J.; SALES, M.; FONSECA, A.; RIBEIRO, J.; NUMATA, I.; COCHRANE, M.; BARBER, C.; ROBERTS, D.; BARLOW, J. Ten-year Landsat classification of deforestation and forest degradation in the brazilian Amazon. **Remote Sensing**, v. 5, n. 11, p. 5493–5513, Oct 2013. ISSN 2072-4292. 16, 46, 49, 50

STONEBRAKER, M.; BECLA, J.; DEWITT, D. J.; LIM, K.-t.; MAIER, D.; RATZESBERGER, O.; ZDONIK, S. B. Requirements for science data bases and SciDB. In: **BIENNIAL CONFRENCE ON INNOVATIVE DATA SYSMTEM RESEARCH, 4., 2009. Proceedings...** [s.n.], 2009. Available from: <http://www-db.cs.wisc.edu/cidr/cidr2009/Paper_26.pdf>. 57

STORY; CONGALTON, R. Accuracy assessment: a user's perspective. **Photogrammetric Engineering and Remote Sensing**, v. 52, n. 3, p. 397–399, 1986. 26

SUN, L.; YANG, X.; JIA, S.; JIA, C.; WANG, Q.; LIU, X.; WEI, J.; ZHOU, X. Satellite data cloud detection using deep learning supported by hyperspectral data. **International Journal of Remote Sensing**, v. 41, n. 4, p. 1349–1371, 2020. 17

TANGE, O. et al. Gnu parallel-the command-line power tool. **The USENIX Magazine**, v. 36, n. 1, p. 42–47, 2011. 10

TEAM, R. **R: A language and environment for statistical computing**. [S.l.]: Vienna, Austria, 2013. 10

TROPEK, R.; SEDLA EK, O.; BECK, J.; KEIL, P.; MUSILOVA, Z.; IMOVA, I.; STORCH, D. Comment on "High-resolution global maps of 21st-century forest cover change". **Science**, v. 344, n. 6187, p. 981–981, May 2014. ISSN 0036-8075, 1095-9203. 1

TYUKAVINA, A.; HANSEN, M. C.; POTAPOV, P. V.; STEHMAN, S. V.; SMITH-RODRIGUEZ, K.; OKPA, C.; AGUILAR, R. Types and rates of forest disturbance in Brazilian Legal Amazon, 2000–2013. **Science Advances**, v. 3, n. 4, 2017. 16

USMAN, S. A.; NITZ, A. H.; HARRY, I. W.; BIWER, C. M.; BROWN, D. A.; CABERO, M.; CAPANO, C. D.; CANTON, T. D.; DENT, T.; FAIRHURST, S.; KEHL, M. S.; KEPPEL, D.; KRISHNAN, B.; LENON, A.; LUNDGREN, A.; NIELSEN, A. B.; PEKOWSKY, L. P.; PFEIFFER, H. P.; SAULSON, P. R.; WEST, M.; WILLIS, J. L. The PyCBC search for gravitational waves from compact binary coalescence. **Classical and Quantum Gravity**, v. 33, n. 21, p. 215004, Nov 2016. ISSN 0264-9381, 1361-6382. 59

VINHAS, L.; RIBEIRO, G.; FERREIRA, K. R.; CAMARA, G. Web services for big Earth observation data. In: **BRAZILIAN SYMPOSIUM ON GEOINFORMATICS (GEOINFO), 17., 2017. Proceedings...** Campos do Jordao, SP, Brazil: INPE, 2017. p. 26–35. 58

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; VAN DER WALT, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VANDERPLAS, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; VAN MULBREGT, P. SciPy 1.0:

fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, n. 3, p. 261–272, Mar 2020. ISSN 1548-7091, 1548-7105. 59

WANG, J.; CHAGNON, F. J. F.; WILLIAMS, E. R.; BETTS, A. K.; RENNO, N. O.; MACHADO, L. A. T.; BISHT, G.; KNOX, R.; BRAS, R. L. Impact of deforestation in the Amazon basin on cloud climatology. **Proceedings of the National Academy of Sciences**, v. 106, n. 10, p. 3670–3674, Mar 2009. ISSN 0027-8424, 1091-6490. 17

WEIER, J.; HERRING, D. **Measuring vegetation (NDVI & EVI)**. 2000. Available from: <https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_1.php>. 41

WILSON, E. H.; SADER, S. A. Detection of forest harvest type using multiple dates of Landsat TM imagery. **Remote Sensing of Environment**, v. 80, n. 3, p. 385 – 396, 2002. ISSN 0034-4257. 41

WOLANIN, A.; CAMPS-VALLS, G.; GOMEZ-CHOVA, L.; MATEO-GARCIA, G.; TOL, C.; ZHANG, Y.; GUANTER, L. Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations. **Remote Sensing of Environment**, 2019. ISSN 00344257. 18

WULDER, M. A.; LOVELAND, T. R.; ROY, D. P.; CRAWFORD, C. J.; MASEK, J. G.; WOODCOCK, C. E.; ALLEN, R. G.; ANDERSON, M. C.; BELWARD, A. S.; COHEN, W. B.; DWYER, J.; ERB, A.; GAO, F.; GRIFFITHS, P.; HELDER, D.; HERMOSILLA, T.; HIPPLE, J. D.; HOSTERT, P.; HUGHES, M. J.; HUNTINGTON, J.; JOHNSON, D. M.; KENNEDY, R.; KILIC, A.; LI, Z.; LYMBURNER, L.; MCCORKEL, J.; PAHLEVAN, N.; SCAMBOS, T. A.; SCHAAF, C.; SCHOTT, J. R.; SHENG, Y.; STOREY, J.; VERMOTE, E.; VOGELMANN, J.; WHITE, J. C.; WYNNE, R. H.; ZHU, Z. Current status of Landsat program, science, and applications. **Remote Sensing of Environment**, v. 225, p. 127–147, May 2019. ISSN 00344257. 6, 38

ZHU, C.; LU, D.; VICTORIA, D.; DUTRA, L. V. Mapping fractional cropland distribution in Mato Grosso, Brazil using time series MODIS enhanced vegetation index and Landsat thematic mapper data. **Remote Sensing**, v. 8, n. 1, p. 223–234, 2015. 21

ZHU, X.; HELMER, E. H. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. **Remote Sensing of Environment**, v. 214, p. 135–153, Sept 2018. ISSN 00344257. 17, 23

ZHU, Z.; WOODCOCK, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. **Remote Sensing of Environment**, v. 118, p. 83–94, 2012. ISSN 00344257. 16, 17, 21

ZUPANC, A. **Improving cloud detection with machine learning**. 2019. Available from: <https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13>. 17, 21, 22, 23, 33