JONATHAN DA ROCHA MIRANDA

SELECTION OF VIRTUAL REMOTE SENSING LIBRARIES AND MACHINE LEARNING TECHNIQUES FOR DIGITAL IMAGE PROCESSING APPLIED TO COFFEE CROP

LAVRAS
MINAS GERAIS – BRASIL

Julho - 2020

JONATHAN DA ROCHA MIRANDA

SELECTION OF VIRTUAL REMOTE SENSING LIBRARIES AND MACHINE LEARNING TECHNIQUES FOR DIGITAL IMAGE PROCESSING APPLIED TO COFFEE CROP

Tese apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, na área de concentração em Engenharia Agrícola, para obtenção do título de Doutor.

Prof. Dr. Marcelo de Carvalho Alves

Orientador

LAVRAS
MINAS GERAIS – BRASIL

Julho – 2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Miranda, Jonathan da Rocha.
Selection of virtual remote sensing libraries and machine learning techniques for digitalimage processing applied to coffee crop / Jonathan da Rocha Miranda. - 2020.
110 p.

Orientador(a): Marcelo de Carvalho Alves.
Coorientador(a): Edson Ampélio Pozza.
Tese (doutorado) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Metadados. 2. Validação de algoritmos. 3. Cafeicultura. I. Alves, Marcelo de Carvalho. II. Pozza, Edson Ampélio. III. Título.

JONATHAN DA ROCHA MIRANDA

SELECTION OF VIRTUAL REMOTE SENSING LIBRARIES AND MACHINE LEARNING TECHNIQUES FOR DIGITAL IMAGE PROCESSING APPLIED TO COFFEE CROP

Tese apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, na área de concentração em Engenharia Agrícola, para obtenção do título de Doutor.

APROVADA em 07 de Julho de 2020

Drª Cristina Rodrigues Nascimento – UFRPE

Drª Luciana Sanches – UFMT

Drª Margarete Marin Lordelo Volpato – EPAMIG

Dr. Gladyston Rodrigues Carvalho – EPAMIG

Prof. Dr. Marcelo de Carvalho Alves

Orientador

LAVRAS
MINAS GERAIS – BRASIL

Julho – 2020

A Deus,
À minha Mãe, Soraia Gomes da Rocha, ao meu Pai, Geraldo Aniceto de Miranda, aos meus Irmãos,Alessandro Gomes da Rocha, Marcos Tadeu Gomes Miranda e Priscila Emanuelle Gomes Miranda e à minha noiva,Gracielle de Brito Sales pelo amor, confiança, companheirismo, respeito e exemplo de vida, sempre fornecendo incentivo e suporte para o meu desenvolvimento pessoal e profissional,
Dedico.

AGRADECIMENTOS

Nesta jornada, muitas pessoas passaram por minha vida e foram fundamentais para esta conquista. Contudo, foi principalmente em Deus que tive o maior apoio para enfrentar cada obstáculo. Por isso, antes de todos, é Ele que devo a minha gratidão, pela vida e pela a chance de concluir mais uma etapa da minha caminhada.

Algumas pessoas aparecem em nossas vidas para nos tornar melhores. Hoje posso dizer que uma dessas pessoas é o meu amigo, companheiro de profissão e de vida, a minha noiva, Gracielle, a quem agradeço por todos os momentos felizes, pela força, pelo carinho e por me confortar nos momentos mais difíceis.

À minha mãe sou eternamente grata, por sempre estar ao meu lado, independentemente das minhas decisões. A ela devo o que sou hoje e dedico esta vitória.

Ao Prof. Marcelo de Carvalho Alves, pela orientação, disposição, paciência, ensinamentos e amizade, que tornaram possível realizar este trabalho.

Ao Prof. Edson Ampélio Pozza, pela coorientação e disposição, que foram essenciais para execução deste projeto.

À equipe que se formou ao longo da caminhada, especialmente aos companheiros Zaqueu, Darliton, Raiza, Pedro, Michel, Rômulo, Felippe, Marcus, Jade, Lucas, Mateus, Gladson, que batalharam de forma inestimável para execução deste projeto.

A Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG), pelo apoio financeiro.

À Universidade Federal de Lavras, sobretudo ao Departamento de Engenharia Agrícola, que contribuíram para esta realização.

Muito obrigado!

GENERAL ABSTRACT

Remote Sensing allows the possibility of monitoring and reasonably estimating the productivity, plant health and mineral nutrition of the coffee tree. By using sensors coupled to satellites it was possible to obtain information about the spectral signature of the coffee crop on a time scale relevant to the monitoring and detection of phenological changes. Surface reflectance can be obtained from a number of remote virtual sensing libraries, each of which can adopt a method of processing the data which may cause divergence in the information. The aim was to evaluate the source of orbital data acquisition and digital image processing for use in coffee growing. Data acquisition source was analyzed for radiometric and geometric differences between SATVeg, AppEEARS and Google Earth Engine platforms comparing a total of 900 sample points distributed in 3 scenes of MOD13Q1 for different dates. Regarding image processing, Machine Learning Random Forest, Naive Bayes and Rede Neural algorithms were evaluated to detect necrosis of coffee tree fruits by comparing the acuractia of the models using Friedman Nemenyi's method. To evaluate whether the Machine Learning algorithms are more effective than the agrometeorological spectral model, estimated productivity by both models using the time series of Landsat images. Based on the results the virtual platforms GEE and AppEEARS can be used with satisfactory accuracy regarding the radiometric values in the condition of being in the sinusoidal projection. Regarding the use of machine learning techniques to detect necrosis in coffee tree fruits, with the Naive Bayes method there were better results in the detection of fruit necrosis through Landsat images. On the yield estimation of coffee tree fruits, with the Random Forest method better estimates were observed in relation to the spectral agrometerological model, being a more indicated method when one intends to estimate the yield at pixel level of Landsat, in the area conditions and availability of images in which the experiment was performed.

Keywords - Metadata, Validation of algorithms, *Colletrochium ssp*, Spectral behaviour, Spectral agrometerological model, Precision agriculture.

RESUMO GERAL

Com o uso do Sensoriamento Remoto há possibilidade de monitorar, estimar com acertos razoáveis a produtividade e o estado nutricional e fitossanitário de cafeeiros em produção. Com sensores acoplados em satélites foi possível obter informações acerca da assinatura espectral da cultura do café numa escala de tempo pertinente ao monitoramento e detecção de mudanças fenológicas. A reflectância da superfície pode ser obtida em diversas bibliotecas virtuais de sensoriamento remoto, podendo cada uma adotar um método de processamento dos dados o que pode causar em divergência na informação. Objetivou-se avaliar a fonte de aquisição de dados orbitais e processamento de imagens digitais para uso na cafeicultura. Para avaliar a fonte de aquisição de dados, foi analisado se existe divergências dos valores radiométricos e geométricos entre as plataformas SATVeg, AppEEARS e Google Earth Engine comparando um total de 900 pontos amostrais distribuídos em 3 cenas do MOD13Q1 para diferentes datas. Em relação ao processamento de imagens, foram avaliados os algoritmos de Machine Learning Random Forest, Naive Bayes e Rede Neural na detecção da necrose dos frutos do cafeeiro comparando a acurácias dos modelos pelo método de Friedman Nemenyi. Para avaliar se os algoritmos de Machine Learning são mais eficazes que o modelo agrometeorologico espectral, estimado a produtividade por ambos modelos utilizando a serie temporal das imagens Landsat. Com base nos resultados as plataformas virtuais GEE e AppEEARS podem ser utilizadas com acurácia satisfatória quanto aos valores radiométricos na condição de estar na projeção sinusoidal. Sobre o uso de técnicas de machine learning para detectar necrose em frutos de cafeeiro, com o método Naive Bayes houve melhores resultados na detecção da necrose de frutos por meio de imagens Landsat. Com relação à estimativa de produtividade de frutos de cafeeiro, com o método Random Forest observaram-se melhores estimativas em relação ao modelo agrometerológico espectral sendo um método mais indicado quando se pretende estimar a produtividade em nível de pixel do Landsat, nas condições de área e disponibilidade de imagens em que o experimento foi realizado.

Palavras chave: Metadados, Validação de algoritmos, *Colletrochium* ssp, variação espectral, Modelo agrometerológico espectral, Agricultura de precisão.

Sumário

1 GENERAL INTRODUCTION

Remote sensing is the science of obtaining information from a target's electromagnetic spectrum through sensors. Based on the interaction of electromagnetic energy from the sun with the Earth's surface, the satellite is able to capture the radiancy emitted by the Earth and thus convert into reflectivity of the surface (HAN; LIU, 2015). Regarding the interpretation of targets, in principle, all can be differentiated according to the reflectance in each range of the electromagnetic spectrum, which can be defined by behavior or spectral signature of the targets. Using this information, one can enter a universe of spectral indices that highlight one target in relation to the other, such as the vegetation indices that highlight the vegetation of other targets (BATRES, 1998).

Data from the spectral signature of the target are relevant for agricultural monitoring due to prior knowledge of typical vegetation behavior, and any change may represent some phytosanitary disturbance (MARTINELLI *et al.*, 2015). Agricultural monitoring using orbital data is being carried out with a focus on estimating spatial variability in productivity, diseases and identification of coffee crops using digital images (BERNARDES *et al.*, 2012; CHEMURA; MUTANGA; ODINDI; *et al.*, 2018; CHEMURA; MUTANGA; SIBANDA; *et al.*, 2018; CHEMURA; MUTANGA; DUBE, 2017; MOREIRA; ADAMI; RUDORFF, 2004).

Remote sensing data acquisition sources are available on virtual platforms where various data options such as the time series of vegetation indexes, processed orbital images or raw file for multi-platform orbitals are available. Using the SATVeg (Vegetation Temporal Analysis System), it is possible to obtain the NDVI (Normalized Difference Vegetation Index) time series and the EVI (Enhanced Vegetation Index) from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor. Google Earth Engine (GEE) through java or python programming can access and manipulate information from the Google Geographic Database that contains multi-platform orbital images. The Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) constitutes a MODIS sensor image database that allows the download of images in blocks at user-defined time intervals.

Virtual platforms are an alternative that would enable the democratization of remote sensing, however, before indicating a platform, tests must be conducted to ensure the reliability of the information provided (DUVAL *et al.*, 2002; KARAMI;

RANGZAN; SABERI, 2013). Methods of image processing such as the configuration of cutting algorithms, reprojection of coordinates among other processes, in case of differences between platforms may occur in the acquisition of different radiometric values. Metadata of the platform image processing when not available are inaccurate regarding the algorithms used and their configuration. In this respect, a research that evaluates the divergences between the platforms are fundamental for the recommendation to users regarding the source of remote sensing information.

Coffee is the most consumed beverage in the world, which makes it a key player in the economic sector. In the stock market, coffee is the agricultural commodity with the highest volume of trading and its contracts are traded on the Commodity and Futures Exchange. Brazil, as the largest coffee producer in the world, drives a long economic chain reaching thousands of people, which in turn favors the generation of employment (CONAB, 2020).

Machine Learning algorithms are promising for image processing applied in coffee farming due to their high adaptability to the data set (SINGH *et al.*, 2016). Machine Learning starts from the assumption that from a database you are able to learn a pattern of behavior and thus extrapolate to various applications. There are a multitude of algorithms applied to filter, classify or even estimate a certain variable (CARRIJO *et al.*, 2017; CHEMURA; MUTANGA; SIBANDA; *et al.*, 2018; SARMIENTO *et al.*, 2014). Depending on the trained database, each algorithm can be better adapted to another. In this sense, it is necessary to establish criteria for choosing algorithms and thus define the one that best behaves with the trained base (MIRANDA *et al.*, 2020).

Combining Machine Learning methods in orbital images can facilitate crop estimation and the detection of diseases in the coffee tree. Mapping the spatial variability of disease can guide producers in places with potential presence of disease and thus carry out a precise, agile and less costly control (MARTINELLI *et al.*, 2015). In relation to yield, in possession of a crop estimate map, the farmer can take measures to manage and conduct the crop as well as plan for the acquisition of inputs, machinery and credit in the market from the perspective of future sales (ROSA *et al.*, 2010).

Spatial variability maps can be obtained by processing orbital images. The radiance of the surface that is captured by the orbital platform acclopated sensor system, when converted into reflectance, collects information pertinent to the monitoring of diseases in the coffee tree (MOREIRA; ADAMI; RUDORFF, 2004). Among the

diseases of the coffee tree, the necrosis of the fruits caused by Colletrochium sp. have symptoms that directly attack the leaves, such as the drought of the hands and the chlorosis of the leaves (FAZUOLI, 2001). Such aspects can influence the spectral behaviour of the coffee tree and thus can be perceptible in orbital images. Therefore, they can be investigated for the Machine Learning method capable of mapping the incidence of fruit necrosis in coffee plantations.

Yield estimation can also be estimated by satellite images, the agrometeorological spectral model being the most widespread in the academic community, however, it is a method that requires spatialized meteorological data in the region where the model is intended to be applied (ALMEIDA, 2013). For small areas these data are punctual and do not have the spatial variability desired for the application of this method. The use of orbital images is part of the agrometerological spectral model in the composition of spectral indices, in this sense, has the tendency to converge in some model of crop estimation with the use of Machine Learning. Once a method of estimation is established at the Landsat pixel level, the crop estimation model can be applied by small to large coffee producers.

1.1 Hypothesis

Options for virtual sensing platforms (SATVeg, GEE and AppEEARS) and digital image processing can be used to define criteria for selecting orbital data acquisition and digital image processing methods best suited to coffee crops.

1.2 Objective

Review the source of orbital data acquisition (SATVeg, GEE and AppEEARS) and digital image processing for use in coffee growing

1.3 Specific objectives

- Assess the reliability of SATVeg, Google Earth Engine and AppEEARS virtual remote sensing platforms for radiometric values and geometric position.

- To evaluate Machine Learning algorithms that are most appropriate for the detection of necrosis of coffee tree fruits in Landsat images.
- To evaluate if with the use of Machine Learning it is possible to obtain better results in yield estimation in relation to the agrometeorological spectral model traditionally used in coffee growing.

1.4 Organization of the thesis

Thesis was organized according to the rules of the Federal University of Lavras for the article structure model.

General introduction of the research was presented in the first part of the thesis. Presentation of this topic was important for a junction of the articles demonstrating that it is in fact a unique and broad research of remote sensing in coffee growing. Papers on the second part of the thesis were presented, which were elaborated in the sequence.

Article 1 entitled **Use of machine learning in digital processing of satellite images applied to coffee crop** provided a state-of-the-art approach on the use of orbital images in coffee growing, demonstrating the evolution of methods used over time, the insertion of machine learning techniques, a survey of algorithms used and evaluation metrics. This article was considered in the literature review topic.

Article 2 entitled **Geometric and radiometric evaluation of remote sensing information in virtual platforms** addressed the evaluation of the reliability of remote sensing virtual platforms. In this research, the EVI radiometric values of MOD13Q1 from Google Earth Engine, SATVeg and AppEEARS platforms were compared with native Earth Explore images. The geographical position between the platforms was evaluated under the condition of maintaining the MODIS sinusoidal projection and reprojecting to geographical coordinates. Results were used to establish criteria for image acquisition with radiometric and geometric errors of satisfactory magnitude to ensure the reliability of future work methodologies.

Article 3 entitled **Detection of coffee berry necrosis by digital image processing of Landsat 8 OLI satellite imagery** is about Machine Learning tests on Landsat image processing for the detection of coffee berry necrosis. In this work the Naive Bayes, Random Forest and MultiLayer Perceptron algorithms were tested in cross validation on 10 parcels repeated 30 times per generating seed. Analysis was

performed for three atmospheric correction models, ATCOR, DOS and 6SV. In this context, about 900 combinations of Machine Learning algorithms were evaluated to define which obtained the best performance in accuracy.

Article 4 entitled **Remote evaluation of the coffee yield by machine learning techniques and spectral agro-meteorological model** made a comparison of the productivity estimation between the spectral agro-meteorological method and Machine Learning. Agrometeorological spectral model was the most used method to estimate productivity using orbital data. Most of this work was done to estimate productivity for low scale, such as the southern region of Minas Gerais or São Paulo. Research was developed to adapt the model to larger pixel level Landsat. Machine Learning Random Forest algorithm was used to compare if there will be a gain in productivity estimation in relation to the agrometeorological spectral model.

REFERENCES

ALMEIDA, Thomé Simpliciano. Modelagem agrometeorológica-espectral para estimativa da produtividade de cafeeiros para areas irrigadas do noroeste de Minas Gerais. 2013. *Dissertação*. 63 f. Universidade Federal de Viçosa, 2013.

BATRES, Vera Beatriz Köhler. Sensoriamento Remoto no estudo da vegetação breve Revisão. *Boletim de Geografia*, v. 16, n. 1, p. 107–118, 1998.

BERNARDES, Tiago *et al.* Monitoring biennial bearing effect on coffee yield using MODIS remote sensing imagery. *International Geoscience and Remote Sensing Symposium (IGARSS)*, v. 4, n. 9, p. 3760–3763, 2012.

CARRIJO, Gabriel L.A. *et al.* Automatic detection of fruits in coffee crops from aerial images. 2017, [S.l.]: IEEE, 2017. p. 1–6.

CHEMURA, Abel; MUTANGA, Onisimo; SIBANDA, Mbulisi; *et al.* Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. *Tropical Plant Pathology*, v. 43, n. 2, p. 117–127, 2018.

CHEMURA, Abel; MUTANGA, Onisimo; ODINDI, John; *et al.* Mapping spatial variability of foliar nitrogen in coffee (Coffea arabica L.) plantations with multispectral Sentinel-2 MSI data. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 138, n. 1, p. 1–11, 2018.

CHEMURA, Abel; MUTANGA, Onisimo; DUBE, Timothy. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. *Precision Agriculture*, v. 18, n. 5, p. 859–881, 2017.

CONAB. *Acompanhamento da safra brasileira 2019/2020*. Disponível em: <https://www.conab.gov.br/info-agro/safras/cafe>. Acesso em: 2 jun. 2020.

DA ROSA, Viviane Gomes Cardoso *et al.* Estimativa da produtividade de café com base em um modelo agrometeorologico-espectral. *Pesquisa Agropecuaria Brasileira*, v. 45, n. 12, p. 1478–1488, 2010.

DUVAL, Erik *et al.* Metadata principles and practicalities. *D-Lib Magazine*, v. 8, n. 4, p. 1082–9873, 2002.

FAZUOLI, L C. O complexo Colletotrichum do cafeeiro. *Campinas: Instituto Agronômico, Boletim Técnico IAC*, 2001.

HAN, Ling; LIU, Dawei. A remote sensing image fusion method based on wavelet transform. *Information Technology and Applications - Proceedings of the 2014 International Conference on Information technology and Applications, ITA 2014*, v. 20, n. 3, p. 361–364, 2015.

KARAMI, Mojtaba; RANGZAN, Kazem; SABERI, Azim. Using GIS servers and interactive maps in spectral data sharing and administration: Case study of Ahvaz Spectral Geodatabase Platform (ASGP). *Computers and Geosciences*, v. 60, p. 23–33, 2013.

MARTINELLI, Federico *et al.* Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development*, v. 35, n. 1, p. 1–25, 2015.

MIRANDA, Jonathan da Rocha *et al.* Detection of coffee berry necrosis by digital image processing of landsat 8 oli satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, v. 85, n. 3, p. 101983, 2020.

MOREIRA, Mauricio Alves; ADAMI, Marcos; RUDORFF, Bernardo Friedrich Theodor. Spectral and temporal behavior analysis of coffee crop in Landsat images. *Pesquisa Agropecuaria Brasileira*, v. 39, n. 3, p. 223–231, 2004.

SARMIENTO, Christiany Mattioli *et al.* Comparação de classificadores supervisionados na discriminação de áreas cafeeiras em Campos Gerais - Minas Gerais. *Coffee Science*, v. 9, n. 4, p. 546–557, 2014.

SINGH, Arti *et al.* Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*, v. 21, n. 2, p. 110–124, 2016.

2 PART TWO – PAPERS

Papers presented in this part were written with the specific standards of the journals chosen for submission and publication.

The first article entitled **Use of machine learning in digital processing of satellite images applied to coffee crop**, is currently being processed by CAB Reviews - Perspectives in agriculture, veterinary Science, nutrition and natural resouces, ISSN: 1749-8848.

The second article entitled **Geometric and radiometric evaluation of remote sensing information in virtual platforms**, is currently under processing of the journal "Computers & Geosciences", ISSN: 0098-3004.

The third article entitled **Detection of coffee berry necrosis by digital image processing of Landsat 8 OLI satellite imagery** was published in volume 85, year 2020 of the journal "International Journal of Applied Earth Observation and Geoinformation", ISSN 0303-2434, https://doi.org/10.1016/j.jag.2019.101983.

The fourth article entitled **Remote evaluation of the coffee yield by machine learning techniques and spectral agro-meteorological model**, was submitted to the journal "Precision Agriculture", ISSN 1573-1618.

PAPER 1 - THE USE OF MACHINE LEARNING IN DIGITAL PROCESSING OF SATELLITE IMAGES APPLIED TO COFFEE CROP

Journal standards CAB *Reviews – Perspectives in agriculture, veterinary science, nutrition and natural resouces*, ISSN: 1749-8848.

Preliminary version

Jonathan da Rocha Miranda and Marcelo de Carvalho Alves

Address: Department of Engineering, Federal University of Lavras. Campus Universitário, PO Box 3037, ZIP code 37200-000, Lavras, Brazil.

Corresponding author: Marcelo de Carvalho Alves **Email:** marcelo.alves@deg.ufla.br

Abstract: Remote Sensing can be used to monitor and estimate, with reasonable correct answers, the yield, plant health and coffee nutrition. Satellite-coupled sensors can obtain information about the spectral signature of the crop, on a time scale, in order to monitor and detect phenological changes. However, the accumulation of data obtained by orbital sensors make it difficult to understand the relationship between the aspects of coffee. Thus, machine learning can perform data mining and meet the spectral signature patterns that constitute coffee behavior. This literature review sought the survey of research that used machine learning tools applied in digital image processing from satellites for coffee crop monitoring.

Keywords: Remote sensing, mapping crop, Agriculture, Data mining, Landsat, Sentinel- 2, MODIS.

Review methodology: This study was conducted to evaluate the use of machine learning in digital processing of satellite images applied to coffee crop. We search the following databases: Multidisciplinary Digital Publishing Institute (MDPI), Wiley Online Library, Taylor & Francis, Scopus (Elsevier), Springer Science, Scientific Electronic Library Online (SciELO), Institute of Electrical and Electronics Engineers (IEES Xplore), Ingenta Connect, Hindawi Publishing Corporation, American Society of Agricultural and Biological Engineers (ASABE) and Google Scholar. The articles were organized on the Mendeley platform in order of topics: “Data acquisition source for the sensing applied in coffee crop”, “Machine learning in orbital image processing” and “Machine learning applications in processing satellite images in the coffee growing

area.” In addition, we used the references of the articles obtained by the aforementioned method to check additional relevant material.

Introduction

Brazil is the largest producer and exporter of coffee and the second largest consumer of this beverage in the world. The coffee culture is not only a relevant source of income for hundreds of municipalities, but it is also important in the sector for the creation of jobs in national agriculture. About 300 thousand producers, distributed in the states of Minas Gerais, São Paulo, Espírito Santo, Bahia, Rondônia, Paraná, Rio de Janeiro, Goiás, Mato Grosso, Amazonas and Pará are involved in this sector [1].

The monitoring of coffee areas, which is of vital importance to the Brazilian economy, is able to obtain as much information as possible about aspects of plant health, planted area and crop estimates, which are essential for the management planning of this commodity [2]. However, much of this information requires financial investment and labor, which makes it difficult to implement [3]. The Brazilian National Supply Company (CONAB) and the Brazilian Institute of Geography and Statistics (IBGE) obtain official information on coffee cultivation, such as area planted and productivity using municipal information obtained by applying standard questionnaires to producers, cooperatives and representatives of public and private agencies [4].

Economic and reliable monitoring methods are crucial due to the high level of long-term investment in the coffee sector [5]. Considering this, the use of orbital Remote Sensing (RS) techniques can represent a significant advance for coffee data researches, mainly aiming at complementing the techniques currently used [6]. The advantage of this system is mainly due to its continuity and spatial extension, culture spectral information and low cost, depending on the type of satellite. By analyzing the time series, the orbital RS is generally more economical and offers an archive of data already acquired through orbital sensors [7].

RS is defined as the analysis of the interactions between electromagnetic radiation and the various substances on Earth captured by a set of sensors and data processing equipment, installed on platforms, with its application in the study of events, phenomena and processes [8]. Based on the reflectance of the Earth’s surface targets, it is possible to enter a universe of vegetation indices that are able to build seasonal and

temporal profiles of vegetation activities, being also possible to detect green peaks, phenological changes of leaves and periods of senescence [9]. RS is an indirect assessment technique, capable of monitoring distance conditions and assessing spatial conditions extension and patterns of vegetation characteristics [10].

However, the use of SR in coffee growing encounters difficulties in establishing assessment methods due to the diversity of management, such as crops under different spacing conditions, plant variety and other factors, such as climate, terrain and illumination [11,12]. Coffee crop is not a seasonal activity and, therefore, in the same region there may be coffee plantations of different ages, which also affect the spectral patterns observed [13]. There are still extensive research projects aimed at using orbital images to monitor coffee growing, even under these difficulties.

Data acquisition source for remote sensing applied in coffee growing

During the selection and acquisition of orbital data, the satellite's characteristics should be oriented as to its periodicity (temporal resolution), pixel area extension (spatial resolution), available spectral ranges (spectral resolution) and radiometric value amplitude (radiometric resolution). Different satellite options are on the market and can meet a large part of the desired specificity; however, only the free platforms will be described, including its applicability in coffee crop.

Land Remote Sensing Satellite (Landsat) is the oldest satellite program in operation since 1972, being the most robust in orbital imaging history [14]. The image pixels cover a terrain of 30 by 30 meters, with 8 bands in the spectral range of 430 to 12510 nm, whose digital values are 8 bits for Landsat 5 TM and 7 ETM+ and Landsat 8 OLI 12 bits (Table 1). Only Landsat 7 ETM+ and Landsat 8 OLI satellites are currently in operation, given that both have a time interval for image acquisition of 16 days, but when the data periodicity is, it can reach up to 8 days [15]. There are major applications in coffee growing with this satellite due to its collection of historical images, which, in turn, has the application in monitoring the coffee area over time. In this area, references highlight works of coffee mapping by age [16], leaf rust detection [17], spatial variability of water consumption [18] and shaded coffee identification [19].

Moderate Resolution Imaging Spectroradiometer – MODIS is a sensor operating on board the Terra and Aqua satellites, with 2,330 km wide detections, covering the

entire surface of the Earth in the range of one to two days. This sensor provides information on aspects of the Earth System since March 2000 [20]. The MODIS sensor covers 36 spectral bands, from 405 to 14.385 nm, and acquires data in three spatial resolutions – 250, 500 and 1000 meters [21]. They are increasingly attractive because of the information and applications that can be worked with, since they have good spectral and temporal resolution characteristics [22,23]. Because of the temporal continuity of MODIS images, there are works applied with time series techniques in coffee growing as an estimate of the area planted [24] and yield forecast [25–29].

Sentinel-2, launched in 2015 by the European Copernicus program, has a 13-band high-resolution spectral imaging covering the range of 430 to 1375 nm, with revisiting time that combined with the Sentinel-2A and -2B satellites reach within five days of revisiting, having a high spatial resolution of 10 meters [30,31]. When compared to Landsat, the Sentinel-2 incorporates three more specifically centralized bands between 705, 740 and 783 nm, which allows for a broad constitution of vegetation indices working in this spectral band [32]. In coffee growing, papers used to assess nitrogen content stand out [33], happening the same with foliar chlorophyll [34] and rust detection [35].

Table 1. Specifications of the satellites most commonly used in remote sensing in coffee growing and practical application of the products made available.

Satellites	Specifications	Application in coffee crops
Terra / Aqua MODIS	250, 500 and 1000 m pixel, 36 spectral bands, image every 1 - 30 days	Monitoring Biennial Bearing Effect [2] Phenological characterization [36] Spectral Agrometeorological model [28] Crop mapping [37]
Landsat 8 OLI	15 and 30 m pixel, 11 spectral bands, images every 16 days	Crop mapping [38] Spectral mixture analysis [39] Identification of biotic and abiotic variables [40] Monitoring of bacterial blight [41]
Sentinel 2 MSI	10 and 20 m pixel, 13 spectral bands, images every 10 days	Mapping [42] Separability leaf rust infection levels [35] Mapping [43] Leaf chlorophyll content [34]

History of orbital data use in coffee growing

From 1985 to 1995, the application of orbital data in coffee growing began with the launch of the Landsat family. The potential use of images was at first in the detection, identification and mapping of the culture [44–48]. Normally, simple image interpretation techniques were used, such as recognition of pattern, texture and tonality and evaluation of spectral behavior to distinguish one culture from another.

By having greater knowledge of remote sensing technologies, other aspects of coffee growing were evaluated, such as height, age, cultivar, plant radius and row spacing [49]. The techniques used were a correlation analysis between biotic attributes and surface reflectance for the Landsat 5 TM satellite.

From 1995 to 2005, the number of works focused on coffee growing was greater as a result of the greater availability of images from orbital platforms such as Terra, Aqua, Landsat 5 TM and 7 ETM+. Research in this period focused on mapping and identifying coffee crops. However, there was an improvement in techniques, such as the use of cluster identification algorithms as the maximum likelihood [50,51] and Mahalanobis Distance [52]. Regression was used in research to find a pattern of spectral behavior as a function of biotic factors of the crop, being an estimation of biomass by multiple regression [53].

Time series research has been implemented in coffee growing to assess spectral variability due to the Landsat image collection available over time [54]. A paper that was published in 2005 aimed at demonstrating the potential of Unmanned Aerial Vehicles (UAVs) for monitoring the ripening of coffee berries using neural network algorithms, with this being one of the first studies using UAVs and machine learning deployment [55].

From 2005 to 2015, there was an increase in research using multi-platform orbital images, mainly the incorporation of MODIS sensor images for crop monitoring focusing on yield estimation [25–29]. Based on MODIS sensor images, it was possible to cover the entire Brazilian territory for the same date. The mapping research of coffee plantations started to address extensive areas, such as the whole state of Minas Gerais [56], which is the largest national producer of coffee. Landsat images were still widely used for mapping the coffee tree, and the most common algorithms had a maximum likelihood [57,58], of the Iterative Self-Organizing kind (ISODATA) [38].

Research on the correlation between GeoEye-1 images and biomass has been developed for biotic variables in coffee [59], with multiple regression on Landsat images to estimate height, density, productivity, vegetative vigor [60] and the phenology of coffee with the use of MODIS sensor images [36,61].

Machine Learning was used in a few works during this period, and it was mainly employed in the identification of coffee crops with the use of algorithms, Naïve Bayes [62], Support Vector Machine [6] and Neural Network [63].

In the last 5 years of published papers, a migration of orbital image processing, with techniques such as supervised maximum likelihood classification, linear or multiple regression and correlation analysis to machine learning has been noted.

Most papers address identification of coffee crops with Decision Tree algorithms [64], Support Vector Machine [64] and Neural Network [65]. Machine learning allows the understanding of patterns often unidentified by analysis of classical statistics. In this sense, the association of satellite images with the biotic variables of the coffee tree has been adopted with the use of ML for the detection of necrosed fruit [66], rust severity prediction [67] and leaf chlorophyll prediction [34].

Machine learning in orbital image processing

Machine Learning (ML) is an area of Artificial Intelligence whose objective is the development of computational techniques for learning, as well as the construction of systems capable of automatically acquiring knowledge. The learning system is a computer program that makes decisions based on the experience accumulated through the successful resolution of previous problems. [68]. ML refers to a group of computational modeling approaches that can learn patterns from a data set to make automatic decisions without programming explicit rules [69].

The execution of the machine learning techniques follows the sequence, database constitution, pre-processing, dimensional reduction, computational model choice and model validation (Figure 1).

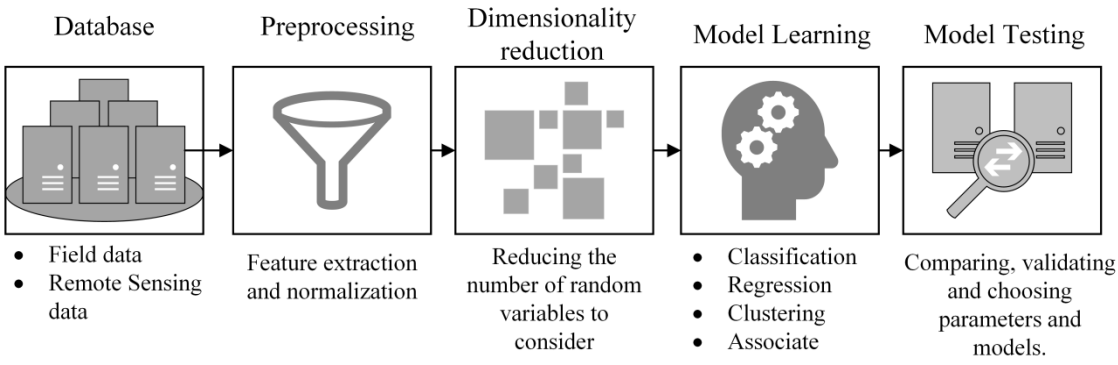

Figure 1. Process steps for implementing remote sensing machine learning

The majority of ML processes make no distinction whether the data are qualitative or quantitative in nature to constitute the independent variables. The database for ML only requires an organization of the data to facilitate the algorithm in linking the dependent variables to the independent variables.

Preprocessing includes the step of identifying errors, recognizing the nature of the errors and using the most appropriate methods to resolve them. Outliers refer to a mechanism for detecting anomalies in a series of data. If a data has a discrepant value regarding the set of samples, it can raise suspicion that it is an error, a noise or even an anomaly [70,71]. Normalization is a stage of preprocessing that performs the transformation of data in order to standardize it in specific ranges, such as 0 and 1 [72]. The normalization is an important tool for ML, as it facilitates the training of algorithms, which tends to reduce the risk of the non-converging estimator [73].

The dimensional reduction in ML allows the elimination of subsets of attributes. The high number of data set dimensions increases the complexity of the techniques and degrades the performance of the data mining algorithms [74]. Data reduction techniques aim to represent a set with the smallest possible size without loss in the characteristics of the sample set [75].

Data reduction techniques in remote sensing have wide application in hyperspectral imaging, and Principal Component Analysis (PCA) is the most employed due to its low complexity and absence of parameters [76]. The disadvantage of the PCA, however, is the high operational cost and low effectiveness when it comes to hyperspectral data [77]. New methods have been used to reduce data in remote sensing, such as Novel Folded-PCA [77], Kernel Principal Component Analysis (KPCA) [76] and wavelet packet transform [78].

ML algorithms are more efficient than classical statistical analysis because they are able to model complex signatures and do not make assumptions about data distribution, being considered as non-parametric analyses [79]. Through this type of applicability, the ML classification has become one of the main focus of the literature on remote sensing [35,79–82]. These methods generally tend to produce higher accuracy compared to traditional parametric classifiers in a wide range of research, especially for complex data, with many predictor variables [82–85].

Different ML algorithms can be applied for data classification and regression, being that in each model different learning techniques are applied, which can achieve better or worse performance depending on the specific problem [86]. In data mining, various algorithms can be used to classify images or to define patterns considering spatial, spectral and temporal features. Regarding the spatial pattern, algorithms such as k-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) are the most recommended.

k-Nearest Neighbor algorithm [87] is one of the simplest machine learning algorithms. It offers two advantages, as it uses a non-parametric approach and allows the use of robust and noisy training data. The algorithm is, therefore, becoming one of the most popular methods applied to forest inventory using remote sensing data in large areas [88–91].

Random Forest [92] is a combination of the prediction of decision trees that are formed by the values of a data set sampled by bootstrap. The criteria for rule ramification are defined by entropy if the classifier and mean absolute error are used in the regression. The classification is made by the percentage of hits that most trees determine. Each activation of the trees produces a generalization error limit value, thus avoiding model overfeeding. Random Forest has been more applied with the use of remote sensing to age identification [5], nitrogen in leaves [33] and crop mapping [64].

Bayesian learning processes are based on the assumption that the values given are regulated by probability distributions. This algorithm is used to predict the probability of an object relevance to a particular class. Bayesian algorithms, such as Naïve Bayes, have a performance comparable to artificial neural networks and decision trees, depending on the problem, having even higher processing speed [93]. This algorithm was used by Zhang *et al.* [94] for pattern recognition to identify varieties of coffee by means of medium infrared spectroscopy, having obtained a precision of 73%.

Mukahema *et al.* [62], when mapping areas of coffee in Rwanda, in QuickBird images, obtained an overall accuracy of 87% with the use of Bayesian networks.

Vector Support Machine [95] is a linear binary classifier that assigns a given test sample to a class of one of two possible labels [81]. This algorithm is particularly attractive in remote sensing applications due to its ability of successfully processing short sets of training data, generally producing a higher classification accuracy than traditional methods [96].

Multilayer Perceptron (MLP) [97] consists of a neural network, in a set of activations, that propagate through a network structure, activated by input data and resulting in an output activation pattern [98]. The learning process of MLP consists of presenting the training data set, and to the extent that there are classification errors. Iterative weights are adjusted and returned, influencing the weights in order to minimize errors in the next iterations. MLP was used to estimate absolute percentages of fruit maturity in coffee. By using spectral images from remotely piloted aircraft research cameras (RPA), the authors obtained a correlation of 0.78 between the NDVI spectral index and the degree of fruit maturation [55].

Cross-validation is generally the most used for ML evaluation [99]. The advantage of this technique is that the entire set of samples is used to evaluate the algorithms and is, therefore, indicated when not stopping large samples. The method consists of randomly dividing the set of training and testing into k-fold. Subsequently, all the training is performed, and the error is calculated using a fold, and the values contained in the fold are returned to the set of samples and a new fold is selected. Then, the entire process is repeated until all the folds have been used [100]. Although the advantages of using cross-validation should be observed, the computational cost of processing is relatively high [99].

Machine learning applications in the digital processing of satellite images in coffee growing

Coffee canopy spectral reflectance can be influenced by humidity, architecture, canopy size, topography, planting density, spacing, cultivar, age, crop consortium and soil fertility [54]. Therefore, it is not just a question of interpreting images to evaluate the conditions of coffee, but, instead, the need for collateral information to understand

why different spectral signatures are necessary for a small variation in space [18]. However, orbital images, with the support of ML algorithms, were able to obtain patterns of behavior and estimate the biophysical aspects of the crop, such as the nitrogen content in leaves [33], incidence and severity of bacterial disease [40], coffee age [16] and plant height [49].

When applying machine learning techniques directly to Landsat 8 OLI images, Miranda *et al.* [66] evaluated the best combination of Naïve Bayes, random forest and multi-layered perceptron algorithms in different models of atmospheric corrections to classify the level of incidence of fruit necrosis in coffee. Naïve Bayes was the best classifier according to the non-parametric analysis of the Friedman and Nemenyi test.

Similarly to the detection of diseases in coffee by orbital sensors, Chemura *et al.* [35,67], when measuring with a portable spectrometer, evaluated for the same spectral range that is adopted for Sentinel-2, severity of rust (*Hemileia vastatrix*) on coffee leaves through the Random Forest. The authors described that the models were satisfactory to identify coffee plants in different levels of rust infection using spectral indices with R^2 of 0.92.

When mapping coffee crops, Mukashema *et al.* [62] developed a Bayesian network model capable of classifying coffee crops into small-scale agroecosystems in Rwanda, Africa. Spectral mixing occurred due to perennial and annual crops near the coffee areas. According to the authors, the combination of spectral data from the QuickBird satellite was effective in mapping coffee fields with an overall accuracy of 50%. Combining information such as the digital elevation model and the forest location map allowed mapping with 87% accuracy. Kelley *et al.* [19] used the Random Forest algorithm to classify shaded coffee by using a set of Landsat 8 OLI images and physiographic components, such as topography and precipitation in northern Nicaragua, having obtained 80% of accuracy.

The implementation of ML techniques is becoming more important in the context of mapping Brazilian coffee. Silveira *et al.* [65] demonstrated that the use of these techniques was a viable alternative in mapping. This study consisted of developing a system for identifying areas cultivated with coffee using Artificial Neural Networks (ANN) with texture as input variables, as indicated by Haralick and Shanmugam [101].

Souza *et al.* [64] reviewed the effectiveness of using different Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms by using different sets of variables for mapping coffee crop with RapidEye images. The results indicated that SVM was the best classifier, with an accuracy of 88%, and the worst results were obtained by RF and NB, with an accuracy of 76 and 82%, respectively. Sarmiento *et al.*[6] aimed to map coffee areas using QuickBird satellite images and the KNN, SVM and Maximum likelihood algorithms. They observed that the maximum likelihood method was superior to other algorithms, with an overall accuracy of about 94% and kappa index of 0.78.

However, due to the distinct analytical behavior of each ML algorithm, it is recommended that more than one algorithm be evaluated for the best certainty of accuracy produced. As stated by Miranda *et al.* [66], the performance of ML algorithms depends on the organization of data assets, and each algorithm can have a better performance in each case.

Table 2. Digital processing of satellite images with Multilayer Perceptron (MLP), Random Forest (RF), Naïve Bayes (NB), Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) applied in coffee crop.

Machine Learning	Imagery	Application in coffee crop
MLP, RF, NB	Landsat 8 OLI	Detection of berry necrosis [66]
RF	Sentinel-2 MSI	Separability of leaf rust infection levels [35]
RF	Sentinel-2 MSI	Prediction of rust severity on leaves [67]
RF	Landsat 8 OLI	Age in the detection of incongruous patches [5]
RF	Sentinel-2 MSI	Leaf chlorophyll content [34]
RF	Sentinel-2 MSI	Mapping spatial variability of foliar nitrogen [33]
ANN	QuickBird	Classification of cultivated areas [62]
RF	Landsat 8 OLI	Map Complex Shade-Grown [19]
ANN	GeoEye-1	Classification of cultivated areas [65]
DT, NB, SVM, KNN	RapidEye	Classification of cultivated areas [64]
KNN, SVM, MaxVer	QuickBird	Classification of cultivated areas [6]
KNN, RF, ANN	Aerial Images	Detection of Fruits [102]
ANN	Aerial Images	Fruits ripeness evaluation [55]

Discussion

Machine learning tools can be considered recent in terms of applications in coffee growing, given that all works mentioned dates from the last 13 years, with a higher concentration of works for the year of 2016 (Figure 2). The Random Forest algorithm was the most widely used in the papers (40%). This algorithm has the advantage over the others because of the power to separately assess the degree of importance of independent variables, and this value is fundamental to understand the spectral behavior of the crop.

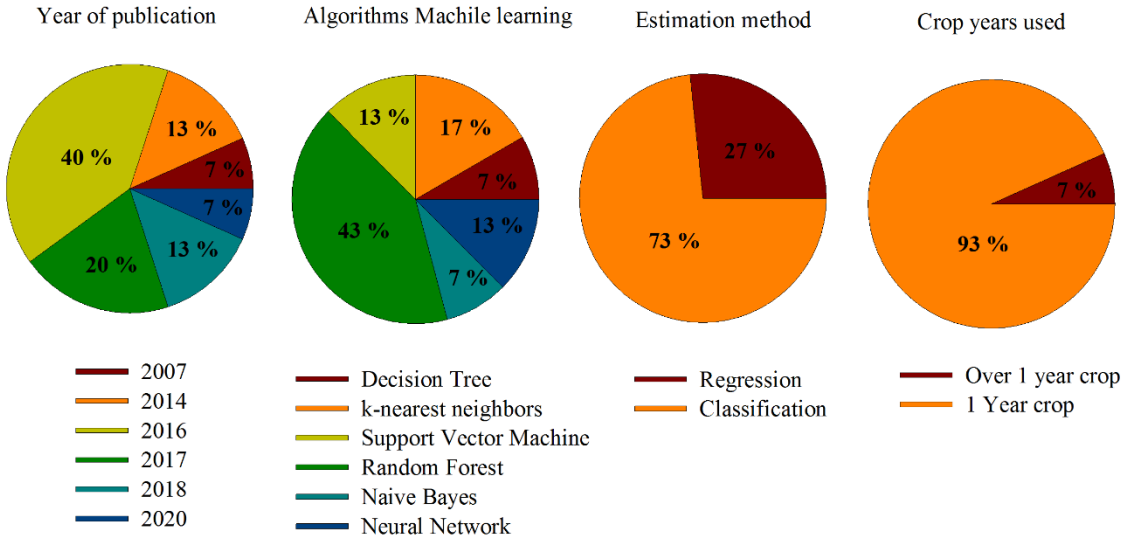

Figure 2. Year of publication, algorithms, methods and number of crop years used in the scientific papers researched [4,6,66,67,94,102,16,17,19,33–35,55,64].

The accuracy of the investigated models was an average of 75% for both classification and regression. A significant aspect of coffee growing is that the dynamics of phenology, over time, can be influenced by the effect of crop bienniality, and, in this case, it requires a study for more than one crop per year. However, 93% of the studies used only one crop per year and, therefore, better precision results can be obtained if more than one year of assessment is done on the effects of crop bienniality.

Literature review was concentrated on scientific papers that used machine learning and remote sensing tools applied to coffee growing, which showed an increased use of the classification (73%) compared to regression (27%) (Figures 2 and

3). The preference for using such classification may be associated with the most used remote sensing image processing software that provides learning of machines for image classification, such as Qgis, ENVI and ArcGis.

Regression application requires one in many cases the domain in programming language, as Scikit-learn module for Python [73], Caret for R [103], Weka for Java, [104] moreover. In regression work, the objective was to estimate biotic and abiotic factors in the culture, which requires the use of georeferenced sample points *in situ* and laboratory analysis. In the classification work, the aim was to map coffee crops where the sample points were less expensive, as they do not require laboratory analysis.

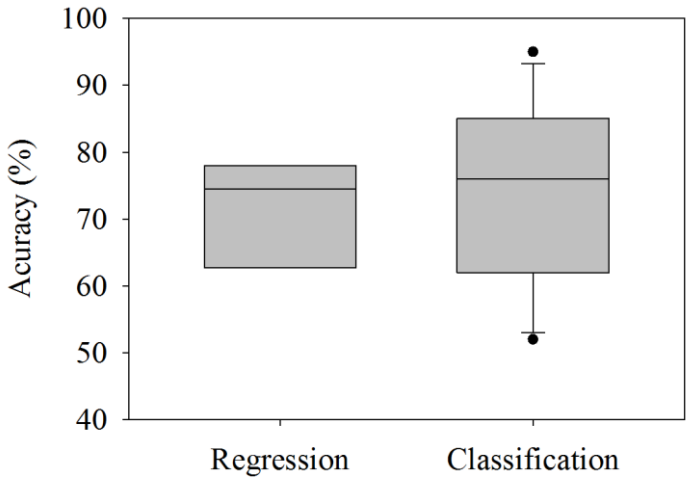

Figure 3. Boxplot of the machine learning algorithms regarding the accuracy of scientific papers researched [4,6,66,67,94,102,16,17,19,33–35,55,64].

Conclusion

Research involving the use of Machine Learning in the processing of digital images applied to coffee culture is largely used for culture mapping. No research was found through the use of these techniques to forecast the coffee harvest. Regarding the detection of diseases, there are few studies addressing this topic, either because it is a more complex approach involving the understanding of variables in coffee culture that can only be obtained *in situ*. Machine learning tools are promising for the processing of satellite images, as the work points to gains in precision and accuracy when compared with approaches that use classical statistics.

Acknowledgements

The authors wish to thank (i) the Department of Agricultural Engineering of the Federal University of Lavras (UFLA) for providing office space and infrastructure to achieve the results obtained in this article and (ii) the Minas Gerais State Research Foundation (FAPEMIG).

References

1. CONAB. Acompanhamento da Safra Brasileira. Vol. 5, Companhia Nacional de Abastecimento. 2019. p. 1–113.
2. Bernardes T, Alves Moreira M, Adami M, Friedrich Theodor Rudorff B. Monitoring biennial bearing effect on coffee yield using MODIS remote sensing imagery. Int Geosci Remote Sens Symp. 2012;4(9):3760–3.
3. Silva S de A, de Queiroz DM, Ferreira WPM, Corrêa PC, Rufino JL dos S. Mapping the potential beverage quality of coffee produced in the Zona da Mata, Minas Gerais, Brazil. J Sci Food Agric. 2016;96(9):3098–108.
4. Kawakubo FS, Pérez Machado RP. Mapping coffee crops in southeastern Brazil using spectral mixture analysis and data mining classification. Int J Remote Sens. 2016;37(14):3414–36.
5. Chemura A, Mutanga O, Dube T. Integrating age in the detection and mapping of incongruous patches in coffee (Coffea arabica) plantations using multi-temporal Landsat 8 NDVI anomalies. Int J Appl earth Obs Geoinf. 2017;57:1–13.
6. Sarmiento CM, Ramirez GM, Coltri PP, Silva LFL, Nassur OAC, Soares JF. Comparison Of Supervised Classifiers In Discrimination Coffee Areas Fields In Campos Gerais-Minas Gerais [comparação De Classificadores Supervisionados Na Discriminação De áreas Cafeeiras Em Campos Gerais-Minas Gerais]. Coffee Sci. 2014;
7. Georgi C, Spengler D, Itzerott S, Kleinschmit B. Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. Precis Agric. 2018;19(4):684–707.
8. Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, *et al.* Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. Ann Assoc Am

Geogr. 2015;105(3):512–30.

9. Batres VBK. Sensoriamento Remoto No Estudo Da Vegetação Breve Revisão [Internet]. Vol. 16, Boletim de Geografia. Parêntese São José dos Campos; 1998. 107–118 p. Available from: http://www.periodicos.uem.br/ojs/index.php/BolGeogr/article/view/12157
10. Martinelli F, Scalenghe R, Davino S, Panno S, Scuderi G, Ruisi P, *et al*. Advanced methods of plant disease detection. A review. Agron Sustain Dev. 2015;35(1):1–25.
11. Friedrich B, Rudorff T. Monitoring Biennial Bearing Effect on Coffee Yield Using MODIS Remote Sensing Imagery. 2012;2492–509.
12. Moreira MA, Rudorff BFT, Barros MA, De Faria VGC, Adami M. Geotecnologies to map coffee fields in the states of minas gerais and são paulo. Eng Agrícola. 2010;30(6):1123–35.
13. Santos JA, Gosselin P-H, Philipp-Foliguet S, Torres R da S, Falao AX. Multiscale classification of remote sensing images. IEEE Trans Geosci Remote Sens. 2012;50(10):3764–75.
14. Tucker CJ, Grant DM, Dykstra JD. NASA's Global Orthorectified Landsat Data Set. Photogramm Eng Remote Sens. 2013;70(3):313–22.
15. USGS. USGS Global Visualization Viewer. US Geol Surv [Internet]. 2012;21:1–11. Available from: url: http://glovis.usgs.gov/index.shtm%5Cnhttp://glovis.usgs.gov/
16. Chemura A, Mutanga O. Developing detailed age-specific thematic maps for coffee (Coffea arabica L.) in heterogeneous agricultural landscapes using random forests applied on Landsat 8 multispectral sensor. Geocarto Int. 2017;32(7):759–76.
17. Katsuhama N, Imai M, Naruse N, Takahashi Y. Discrimination of areas infected with coffee leaf rust using a vegetation index. Remote Sens Lett. 2018;9(12):1186–94.
18. Costa J de O, Coelho RD, Wolff W, José JV, Folegatti MV, Ferraz SF de B. Spatial variability of coffee plant water consumption based on the SEBAL algorithm. Sci Agric. 2019;76(2):93–101.
19. Kelley L, Pitcher L, Bacon C. Using Google Earth Engine to Map Complex Shade-Grown Coffee Landscapes in Northern Nicaragua. Remote Sens.

2018;10(6):952.

20. Vannan SKS, Cook RB, Holladay SK, Olsen LM, Dadi U, Wilson BE. A web-based subsetting service for regional scale MODIS land products. IEEE J Sel Top Appl Earth Obs Remote Sens. 2009;2(4):319–28.
21. Menzel WP, Frey RA, Zhang H, Wylie DP, Moeller CC, Holz RE, *et al.* MODIS global cloud-top pressure and amount estimation: Algorithm description and results. J Appl Meteorol Climatol. 2008;47(4):1175–98.
22. Swain S, Abeysundara S, Hayhoe K, Stoner AMK. Future changes in summer MODIS-based enhanced vegetation index for the South-Central United States. Ecol Inform. 2017;41:64–73.
23. Shi H, Li L, Eamus D, Huete A, Cleverly J, Tian X, *et al.* Assessing the ability of MODIS EVI to estimate terrestrial ecosystem gross primary production of multiple land cover types. Ecol Indic. 2017;72:153–64.
24. Chaves MED, Ferreira E, Dantas AAA. Thresholds definition in MOD13Q1 and VGT-S10 time series for coffee crop area estimation in Triângulo Mineiro/Alto Paranaíba. Theor Appl Eng. 2019;3(2):1–10.
25. Victorino EC, de Carvalho LG, Ferreira DF. Agrometeorological modeling for coffee productivity forecast in the south region of Minas Gerais state. Coffee Sci. 2016;11(2):211–20.
26. Gomes V, Moreira MA. Estimativa da produtividade de café com base em um modelo agrometeorológico - espectral. 2010;(1):1478–88.
27. Almeida TS. Modelagem agrometeorológica-espectral para estimativa da produtividade de cafeeiros para areas irrigadas do noroeste de Minas Gerais. 2013;
28. Rosa VGC, Moreira MA, Rudorff BFT, Adami M. Estimativa da produtividade de café com base em um modelo agrometeorologico-espectral. Pesqui Agropecu Bras. 2010;45(12):1478–88.
29. Monteiro LA, Sentelhas PC. <b>Calibration and testing of an agrometeorological model for the estimation of soybean yields in different Brazilian regions. Acta Sci Agron. 2014;36(3):265.
30. Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, *et al.* Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ. 2012;120:25–36.

31. Immitzer M, Vuolo F, Atzberger C. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sens. 2016;8(3):166.
32. Delegido J, Verrelst J, Alonso L, Moreno J. Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. Sensors. 2011;11(7):7063–81.
33. Chemura A, Mutanga O, Odindi J, Kutywayo D. Mapping spatial variability of foliar nitrogen in coffee (Coffea arabica L.) plantations with multispectral Sentinel-2 MSI data. ISPRS J Photogramm Remote Sens [Internet]. 2018;138:1–11. Available from: https://doi.org/10.1016/j.isprsjprs.2018.02.004
34. Chemura A, Mutanga O, Odindi J. Empirical modeling of leaf chlorophyll content in coffee (Coffea Arabica) plantations with Sentinel-2 MSI data: Effects of spectral settings, spatial resolution, and crop canopy cover. IEEE J Sel Top Appl Earth Obs Remote Sens. 2017;10(12):5541–50.
35. Chemura A, Mutanga O, Dube T. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. Precis Agric [Internet]. 2017;18(5):859–81. Available from: https://doi.org/10.1007/s11119-016-9495-0
36. Júnior AFC, de Carvalho Júnior OA, de Souza Martins É, Guerra AF. Phenological characterization of coffee crop (Coffea arabica L.) from Modis time series. Brazilian J Geophys. 2013;31(4):569–78.
37. Bispo RC, Lamparelli RAC, Rocha J V. Using fraction images derived from modis data for coffee crop mapping. Eng Agrícola. 2014;34(1):102–11.
38. Ortega-Huerta MA, Komar O, Price KP, Ventura HJ. Mapping coffee plantations with Landsat imagery: an example from El Salvador. Int J Remote Sens. 2012;33(1):220–42.
39. Schmitt-Harsh M, Sweeney SP, Evans TP. Classification of coffee-forest landscapes using Landsat TM imagery and spectral mixture analysis. Photogramm Eng Remote Sens. 2013;79(5):457–68.
40. Marin DB, de Carvalho Alves M, Pozza EA, Belan LL, de Oliveira Freitas ML. Multispectral radiometric monitoring of bacterial blight of coffee. Precis Agric. 2019;20(5):959–82.
41. Marin DB, Alves M de C, Pozza EA, Gandia RM, Cortez MLJ, Mattioli MC. Sensoriamento remoto multiespectral na identificação e mapeamento das

variáveis bióticas e abióticas do cafeeiro. Rev Ceres. 2019;66(2):142–53.

42. Erinjery JJ, Singh M, Kent R. Mapping and assessment of vegetation types in the tropical rainforests of the Western Ghats using multispectral Sentinel-2 and SAR Sentinel-1 satellite imagery. Remote Sens Environ. 2018;216:345–54.
43. Clerici N, Valbuena Calderón CA, Posada JM. Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia. J Maps. 2017;13(2):718–26.
44. Batista GT, Tardin AT, Chen SC, Dallemand JF. Avaliação de produtos HRV/SPOT e TM/LANDSAT na discriminação de culturas. Pesqui Agropecuária Bras. 1990;25(3):379–86.
45. Tardin AT, de Assunção GV, Soares JV. Análise preliminar de imagens TM visando a discriminação de café, citrus e cana-de-açúcar na região de Furnas, MG. Pesqui Agropecuária Bras. 1992;27(9):1355–61.
46. KÖHL M, Kushwaha SPS. A four-phase sampling method for assessing standing volume using Landsat-TM-data, aerial photography and field assessments. Commonw For Rev. 1994;35–42.
47. Croome R. The potential for satellite remote sensing to monitor coffee, tea, cocoa and coconut plantings in Papua New Guinea. Experimentation with Landsat MSS and TM Data in the Madang and Goroka areas of PNG. 1989;
48. Veloso MH. Coffe inventory through orbital imagery. Rio Janeiro Inst Bras do Café. 1974;
49. Epiphanio JCN, Leonardi L, Formaggio AR. Relações entre parâmetros culturais e resposta espectral de cafezais. Pesqui Agropecuária Bras. 1994;29(3):439–47.
50. Lu D. Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. Int J Remote Sens. 2005;26(12):2509–25.
51. Lelong CCD, Thong-Chane A. Application of textural analysis on very high resolution panchromatic images to map coffee orchards in Uganda. In: IGARSS 2003 2003 IEEE International Geoscience and Remote Sensing Symposium Proceedings (IEEE Cat No 03CH37477). IEEE; 2003. p. 1007–9.
52. Pedroni L. Improved classification of Landsat Thematic Mapper data using modified prior probabilities in large and complex landscapes. Int J Remote Sens. 2003;24(1):91–113.
53. Lu D, Batistella M, Moran EF, MIRANDA EED. A comparative study of Terra

ASTER, Landsat TM, and SPT HRG data for land cover classification in the Brazilian Amazon. In: Embrapa Territorial-Artigo em anais de congresso (ALICE). In: World multi-conference on systemics, cybernetics and informatics; 2005.

54. Moreira MA, Adami M, Rudorff BFT. Spectral and temporal behavior analysis of coffee crop in Landsat images. Pesqui Agropecuária Bras. 2004;39(3):223–31.
55. Furfaro R, Ganapol BD, Johnson LF, Herwitz SR. Neural network algorithm for coffee ripeness evaluation using airborne images. Appl Eng Agric. 2005;23(3):379–87.
56. Vieira TGC, Alves HMR, Bertoldo MA, Souza VCO de. Geothecnologies in the assessment of land use changes in coffee regions of the state of Minas Gerais in Brazil. 2007;
57. Cordero-Sancho S, Sader SA. Spectral analysis and classification accuracy of coffee crops using Landsat and a topographic-environmental model. Int J Remote Sens. 2007;28(7):1577–93.
58. Martínez-Verduzco GC, Galeana-Pizaña JM, Cruz-Bello GM. Coupling community mapping and supervised classification to discriminate Shade coffee from Natural vegetation. Appl Geogr. 2012;34:1–9.
59. Coltri PP, Zullo J, do Valle Goncalves RR, Romani LAS, Pinto HS. Coffee crop's biomass and carbon stock estimation with usage of high resolution satellites images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2013;6(3):1786–95.
60. Vieira TGC, Alves HMR, Lacerda MPC, Veiga RD, Epiphanio JCN. Crop parameters and spectral response of coffee (Coffea arabica L.) areas within the state of Minas Gerais, Brazil. Coffee Sci. 2007;1(2):111–8.
61. Brunsell NA, Pontes PPB, Lamparelli RAC. Remotely sensed phenology of coffee and its relationship to yield. GIScience Remote Sens. 2009;46(3):289–304.
62. Mukashema A, Veldkamp A, Vrieling A. Automated high resolution mapping of coffee in Rwanda using an expert Bayesian network. Int J Appl Earth Obs Geoinf. 2014;33:331–40.
63. Gomez C, Mangeas M, Petit M, Corbane C, Hamon P, Hamon S, *et al.* Use of high-resolution satellite imagery in an integrated model to predict the distribution

of shade coffee tree hybrid zones. Remote Sens Environ. 2010;114(11):2731–44.

64. Souza CG, Carvalho L, Aguiar P, Arantes TB. Machine learning algorithms and variable of remote sensing for coffee cropping mapping. Bol Ciências Geodésicas. 2016;22(4):751–73.
65. Silveira LS, Valente DSM, Pinto F, Santos FL. Case studies of classification of cultivated areas with coffee by texture descriptors. Coffee Sci. 2016;11(4):502–11.
66. Miranda J da R, de Carvalho Alves M, Pozza EA, Neto HS. Detection of coffee berry necrosis by digital image processing of landsat 8 oli satellite imagery. Int J Appl Earth Obs Geoinf. 2020;85:101983.
67. Chemura A, Mutanga O, Sibanda M, Chidoko P. Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. Trop Plant Pathol. 2018;43(2):117–27.
68. Rezende SO. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda; 2003.
69. Singh A, Ganapathysubramanian B, Singh AK, Sarkar S. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends Plant Sci [Internet]. 2016;21(2):110–24. Available from: http://dx.doi.org/10.1016/j.tplants.2015.10.015
70. Aggarwal CC. Outlier analysis. In: Data mining. Springer; 2015. p. 237–63.
71. Hawkins DM. Identification of outliers. Vol. 11. Springer; 1980.
72. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
73. Fabian P, Michel Vincent, Olivier G, Mathieu B, Peter P, Ron W, *et al.* Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
74. Van Der Maaten L, Postma E, Van den Herik J. Dimensionality reduction: a comparative. J Mach Learn Res. 2009;10(66–71):13.
75. Zhang D, Zhou Z-H, Chen S. Semi-supervised dimensionality reduction. In: Proceedings of the 2007 SIAM International Conference on Data Mining. SIAM; 2007. p. 629–34.
76. Fauvel M, Chanussot J, Benediktsson JA. Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. EURASIP J Adv Signal Process. 2009;2009(1):783194.

77. Zabalza J, Ren J, Yang M, Zhang Y, Wang J, Marshall S, *et al.* Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. ISPRS J Photogramm Remote Sens. 2014;93:112–22.
78. Han L, Liu D. A remote sensing image fusion method based on wavelet transform. In: Information Technology and Applications - Proceedings of the 2014 International Conference on Information technology and Applications, ITA 2014. IEEE; 2015. p. 361–4.
79. Maxwell AE, Warner TA, Fang F. Implementation of machine-learning classification in remote sensing: An applied review. Int J Remote Sens. 2018;39(9):2784–817.
80. Belgiu M, Drăgu L. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens. 2016;114:24–31.
81. Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: A review. ISPRS J Photogramm Remote Sens. 2011;66(3):247–59.
82. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens. 2005;26(1):217–22.
83. Ghimire B, Rogan J, Galiano VR, Panday P, Neeti N. An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. GIScience Remote Sens. 2012;49(5):623–43.
84. Huang C, Davis LS, Townshend JRG. An assessment of support vector machines for land cover classification. Int J Remote Sens. 2002;23(4):725–49.
85. Rogan J, Miller J, Stow D, Franklin J, Levien L, Fischer C. Land-cover change monitoring with classification trees using Landsat TM and ancillary data. Photogramm Eng Remote Sens. 2003;69(7):793–804.
86. Tan P-N, Steinbach M, Kumar V. Introdução ao datamining: mineração de dados. Ciência Moderna; 2009.
87. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern. 1985;(4):580–5.
88. Holmström H. Estimation of single-tree characteristics using the kNN method and plotwise aerial photograph interpretations. For Ecol Manage. 2002;167(1–3):303–14.
89. Katila M, Tomppo E. Selecting estimation parameters for the Finnish multisource

National Forest Inventory. Remote Sens Environ. 2001;76(1):16–32.

90. Tokola T. The influence of field sample data location on growing stock volume estimation in Landsat TM-based forest inventory in eastern Finland. Remote Sens Environ. 2000;74(3):422–31.
91. Tomppo E, Halme M. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Remote Sens Environ. 2004;92(1):1–20.
92. Breiman L. RANDOM FORESTS Leo. Mach Learn. 2001;45(1):1–33.
93. Langley P. Induction of recursive Bayesian classifiers. In: European Conference on Machine Learning. Springer; 1993. p. 153–64.
94. Zhang C, Wang C, Liu F, He Y. Mid-infrared spectroscopy for coffee variety identification: Comparison of pattern recognition methods. J Spectrosc. 2016.
95. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9(3):293–300.
96. Mantero P, Moser G, Serpico SB. Partially supervised classification of remote sensing images through SVM-based probability density estimation. IEEE Trans Geosci Remote Sens. 2005;43(3):559–70.
97. Hinton GE. CONNECTIONIST LEARNING PROCEDURES. Mach Learn [Internet]. 1990 Jan 1 [cited 2019 Apr 2];555–610. Available from: https://www.sciencedirect.com/science/article/pii/B9780080510552500298
98. Pound MP, Atkinson JA, Townsend AJ, Wilson MH, Griffiths M, Jackson AS, *et al.* Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience [Internet]. 2017;6(10):gix083. Available from: http://www.gigasciencejournal.com
99. Blockeel H, Struyf J. Efficient algorithms for decision tree cross-validation. J Mach Learn Res. 2002;3(Dec):621–50.
100. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci (Ny). 2012;191:192–213.
101. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;(6):610–21.
102. Carrijo GLA, Oliveira DE, de Assis GA, Carneiro MG, Guizilini VC, Souza JR. Automatic detection of fruits in coffee crops from aerial images. In: 2017 Latin American Robotics Symposium (LARS) and 2017 Brazilian Symposium on

Robotics (SBR). IEEE; 2017. p. 1–6.

103. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, *et al.* Package ‘caret.’ R J. 2020;

104. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. ACM SIGKDD Explor Newsl. 2009;11(1):10–8.

PAPER 2- GEOMETRIC AND RADIOMETRIC EVALUATION OF REMOTE SENSING INFORMATION IN VIRTUAL PLATFORMS

Journal standards *Computers & Geosciences*, ISSN: 0098-3004

Preliminary version

Jonathan da Rocha Miranda[*1], Marcelo de Carvalho Alves[*2]

Department of Agricultural Engineering, Federal University of Lavras, Campus University,

P.O. Box: 3037, ZIP Code: 37200-000, Lavras, Minas Gerais, Brazil

jhonerocha@estudante.ufla.br

Abstract: With the advent of geographic databases on the web, incorporating multi-data and satellite sensor information, virtual platforms emerged to democratize the access to remote sensing cloud information, with a high computational processing capacity. However, the lack of metadata on information processing referred to virtual platforms leads to questions regarding the quality of that data for different and available applications. Presuming that the search for the same data on different platforms tends not to confront the same solution, the objective of this research was to investigate the geometric and radiometric differences between platforms like Google Earth Engine (GEE), The Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) and Vegetation Temporal Analysis System (SATVeg). Three scenarios (H11V09, H12V10 and H13V11) were evaluated for the EVI image of the MOD13Q1 product, in which they were extracted for the sampling points for the SATVeg, GEE and AppEEARS platforms, being the reference image obtained in the Earth Explore image catalog. The duplication of the pixel was evaluated in a case of two pixels in a row, with the achievement of the same radiometric values in the images, in different dates. The geometry of the image was evaluated by the Euclidean distance, from the center of the Earth Explore reference pixel and the same location in virtual platforms. The radiometric divergence was evaluated by making a comparison, point to point, of the values between Earth Explore and other virtual platforms. Pixels with double radiometric value, as well as geometric and radiometric divergences were found in all platforms when using the geographic projection in the data. However, with the use of sinusoidal projection, there was consistency in all proposed evaluations. Virtual platforms are a great advance in the democratization of remote sensing since they integrate the availability of information in an accessible way without the need of broad knowledge on the subject, with data processing in clouds. For the effective application of these tools, it is necessary to determine the reliability of this information and to make it available as a form of guidance to users and some quality control information if possible.

Keywords: Metadata; EVI; SATVeg; Google Earth Engine; AppEEARS.

1 INTRODUCTION

The emergence of cloud-based geographic information databases, supported by digital image processing tools, made it possible to create virtual remote sensing platforms. The platforms offer a set of data that is relevant to research, so as to require less data processing, allowing researchers to focus more on the search for knowledge and solutions to problems on a global scale (Bailey and Chen, 2011; Quenzer and Friesz, 2015; Zhu et al., 2014). The great advantage of these platforms is the possibility of integrating sectors of other knowledge to acquire and interpret remote sensing data without the necessity of being an expert in the subject.

The Moderate Resolution Imaging Spectroradiometer (MODIS) sensor has as a great advantage due to its continuity and spatial extension, given that spectral information is made available free of charge (Georgi et al., 2018). These data can be easily acquired on virtual remote sensing platforms such as Google Earth Engine (GEE), Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) and Temporal Vegetation Analysis System (SATVeg).

Virtual remote sensing platforms are being widely used in agricultural monitoring research, such as in the observation of storm impacts on crops (Gallo et al., 2019), hydrological assessment (Li et al., 2019), land cover change (Münch et al., 2019), impacts of drought on the environment (Buras et al., 2019), land use changes for sugarcane activity (Melo et al., 2018), impacts of soybean on increased deforestation (Kastens et al., 2017), land cover mapping (Macedo et al., 2018), dynamics of sugarcane production (Antunes et al., 2015), temporal analysis of eucalyptus plantations (Trentin et al., 2018) and delimitation of permanent preservation area (Miranda and Lipp-Nissinen, 2017).

However, it is necessary to establish guidelines as to the reliability of the data in order to be useful for decision-making. Therefore, before implementing projects that use data from virtual platforms, it is necessary to make sure that the information is accurate regarding radiometric and geometric values. In precision agriculture, the reliability of the information regarding the spatial variability of the physiological characteristics of the crop is fundamental to optimize the application of agricultural inputs. The geometric error in this kind of operation can impact on the application of agricultural inputs in undue place, as the radiometric error can lead to a false interpretation of the crop

conditions. In this sense, extreme caution should be taken when relying on data from virtual platforms.

The use of a virtual multiplatform system requires the standardization of images in a single cartographic projection system (Yildirim and Kaya, 2008). Thus, virtual platforms tend to adopt a single projection system for the geographic database. This projection standardization requires the processing of the images, which, in turn, may lead to radiometric divergence if one platform adopts a different method from another. Image projection requires an image resampling, and the procedure may cause the loss or even duplication of the original pixels (Seong et al., 2002).

The reliability of virtual platforms can be consulted based on metadata, which refers, in some way, to the biography of the geographic data, demonstrating the specifications and procedures adopted that originated the data presented (Su et al., 2000; Zurier, 1996). Karami et al. (2013), report that, without the correct description of metadata, it is impossible to use the information for other researchers. The detailed descriptions of the metadata of the virtual platforms allow, therefore, that other users use the tools arranged with precision and with assured quality (Duval et al., 2002).

The lack of metadata that is specific to the internal processing that occurs in these platforms does not guarantee the reliability of the information acquired. Relevant information, such as the methods adopted for image reprojection and resampling, is fundamental to assess whether the data presented can be replicated and, thus, obtain the same results. Due to the inexistence of standardization policy, quality control protocol and interoperability of geospatial data infrastructure, the objective of this scholarly paper was to evaluate the quality of SATVeg, AppEEARs and GEE virtual remote sensing platforms in terms of geographic, radiometric and temporal positioning.

2 MATERIAL AND METHODS

2.1 Descriptions of areas

Three scenarios of the MODIS MOD13Q1 product were selected in different orbits and points (H11V09, H12V10 and H13V11) along the Brazilian territory. The sample points were distributed in 50 points, which were aligned in the east and north directions, being the point of origin the centroid of each scenario (Figure 1).

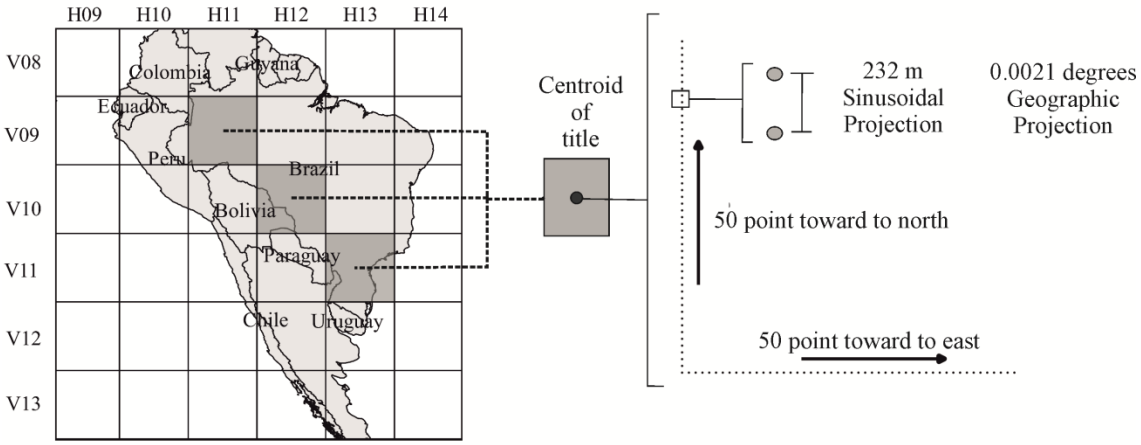

Figure 1. Description of the area with the sample points used to validate the SATVeg, GEE and AppEEARS platforms from the centers of H11V09, H12V10 and H13V11 scenarios.

It was analyzed if the reprojection of coordinates can change the geometry and radiometric values of the images. In this sense, it was used the sinusoidal projection, which corresponds to the native projection of MOD13Q1, being also utilized in the geographical projection. The sinusoidal projection has a metric system of plane coordinates, with the pixel dimension of the MOD13Q1 product at about 232 m, given the fact that this value was considered as the spacing between the sample points. Regarding the geographical projection, the metric system is defined in degrees, with a value for the pixel dimension of the MOD13Q1 product images around 0.00208 degrees, being this value considered as the spacing between the sample points.

GEE and AppEEARS have the availability of extracting the radiometric values in the platform itself and the option of downloading the images. Thus, the radiometric and geometric evaluations were performed by extracting the values extracted in the platform itself and in terms of the values extracted after downloading the images.

2.2 Description of Virtual Remote Sensing Platforms

The GEE Code is a platform that allows the processing of images by programming in Java script or Python language, with the possibility of addition of simple operations or even the most complex processes (Gorelick et al., 2017). Its great advantage is the absence of the need of storing images in a local database, because all

images are in the cloud, which allows more elaborate work to be performed since it does not require the operational cost of software, and the need of hardware is reduced.

AppEEARS offers the MODIS sensor image analysis exploratory services, enabling the interaction and exploration of the requested data and its associated quality information even before downloading them (Maiersperger, 2017; Neeley, 2018; Quenzer and Friesz, 2015). This platform provides the options to obtain time series data from previously entered geographical points or MODIS sensor images. For image access, the requirement is to establish the boundary of the area to be displayed. Besides that, it can be performed manually by drawing the area or inserting a shapefile polygon.

The SATVeg is a virtual platform created by the Brazilian Agricultural Research Company (EMBRAPA) with the purpose of providing access to remote sensing technologies to the most diverse audiences. In this way, it is an interface in which the user has only to indicate the place of interest with a simple click to obtain information of temporal series of NDVI and EVI vegetation indices. Although it does not directly provide an image for analysis, the data obtained can be processed by several filters, such as Savitzky-Golay (Gorry, 1990), Flat Bottom (Wardlow et al., 2006) and the Wavelet transform (Daubechies, 1990). These filters have the function of adjusting the curves, thus remaining sensitive to the seasonal changes in the vegetation without losing its temporal characteristics (Sakamoto et al., 2005). Files can also be inserted in shapefiles, but only for the purpose of supporting the orientation of the desired location.

2.3 Acquisition of radiometric information

Three dates, at five-year intervals, were used to evaluate the platforms (Table 1). The month of August was defined for the analyses because it is a drier period of the year for most of the national territory, which may allow for less interference in the reflectivity of the surface due to the presence of clouds. The choice of the adjusted vegetation index EVI (Justice et al., 1998) was defined as a product originating from the blue, red and infrared bands of the MODIS sensor. In other words, it is an indirect evaluation of three spectral bands of the MOD13Q1 product.

$$EVI = 2.5 * \frac{\rho_{nir} - \rho_{red}}{1 + \rho_{nir} + 6 * \rho_{red} - 7.5 * \rho_{blue}} \quad (1),$$

where: ρnir is the reflectance in the infra-red band; ρred is the reflectance in the red band; ρblue is the reflectance in the blue band.

Table 1. Virtual platforms selected for the study and their respective products under analysis

Virtual platform	Image	Product	Date
SATVeg	MOD13Q1	EVI	August 13, 2009, 2014 and 2019
GEE	MOD13Q1	EVI	August 13, 2009, 2014 and 2019
AppEEARS	MOD13Q1	EVI	August 13, 2009, 2014 and 2019
Image Catalogue Earth Explore	MOD13Q1	EVI	August 13, 2009, 2014 and 2019

The Earth Explore image catalog (URL: https://earthexplorer.usgs.gov) was defined as the reference base for the comparison of data due to the possibility of acquiring native data from the MOD13Q1 images without any type of processing, such as file format conversion or reprojection. The images were used in the sinusoidal projection, which refers to the native projection of the image, that is, not an a priori processing, and in the geographic projection, which were remastered by the method of the nearest neighbor. Therefore, the radiometric values for the sinusoidal and geographical projection were extracted.

In the SATVeg website (URL: https://www.satveg.cnptia.embrapa.br), the vector file of the sample points in shapefile format was inserted in the geographic projection. In each point, it was downloaded the spreadsheet with the EVI values of the MOD13Q1 product for the period from January 2000 to May 2020. The EVI values on the dates under analysis were extracted through a python computational routine capable of accessing all spreadsheets and copying the EVI value specifically on the date under evaluation.

In the GEE (URL: https://code.earthengine.google.com), the shapefile files of the sample points were uploaded in the geographical projection. The extraction of EVI values was acquired through a javascript with the function to extract the values from the MOD13Q1 image, specifically on the desired dates, being subsequently saved in a spreadsheet in csv format. For the acquisition of the images, a script capable of selecting the EVI index of the MOD13Q1 product on the date stipulated for analysis, specifically

in the area extension corresponding to a 50 km radius from the first point in analysis, was elaborated. The images were downloaded in the sinusoidal and geographical projections for all dates in analysis, and the radiometric values were extracted for the sample points in the image projection.

For the AppEEARS platform (URL: https://lpdaacsvc.cr.usgs.gov/appeears), the EVI values were extracted from the images from the MOD13Q1 product. Afterwards, it was inserted in the platform the coordinates of the points and dates of interest that were imported through a file in csv format. Regarding the download of images, a 50 km radius shapefile buffer was imported from the first sample point, and only the images on the dates of interest were selected. These images were acquired in the sinusoidal and geographic projections for all dates under analysis, being the extracted values referred to sample points in the projection regarding the image.

2.4 Geometric Analysis of MOD13Q1 images

The geometric comparison was performed to determine whether the platform images are in the same alignment in terms of their geographical position. This procedure was performed by measuring the Euclidean distance between the points located in the center of the pixel, referring to the first sample point of the images in the sinusoidal and geographic projections.

In the specific case of the SATVeg, in the header of the spreadsheet the geographic coordinates in the center of the pixel were inserted, and, in this case, it has been used the same script in python that extracted the radiometric values, but focusing on the coordinates.

2.5 Analysis of the double radiometric values of the platforms

Duplicate information was considered if the values between two consecutive points are equal on three analysis dates. In case of detection of duplicate points, a new data collection was performed directly on the platforms in order to find out if this duplicity occurred throughout the historical series.

2.6 Analysis of radiometric data consistency

On SATVeg, EVI information is available, which, theoretically, is the same as that found on NASA's MOD13Q1 images. In GEE and AppEEARS, these values are directly searched from the MOD13Q1 images, from the platform's own database. Thus, the evaluation was performed in direct comparison among the EVI values of the SATVeg, GEE and AppEEARS platforms in relation to the Earth Explore images. The analysis occurred point by point and was recorded in terms of whether this value was coincident, displaced or absent in the same spatial location of the points regarding the Earth Explore native images. This procedure was performed for the images in the sinusoidal and geographical projections referred to the points directly extracted from the platform and concerning the images that were downloaded.

In summary, amongst the proposed methodological procedures, the SATVeg, GEE and AppEEARS platforms were analyzed for the possibility of duplicate pixels and divergences regarding the native images from the Earth Explore image catalog, as specified in Figure 2.

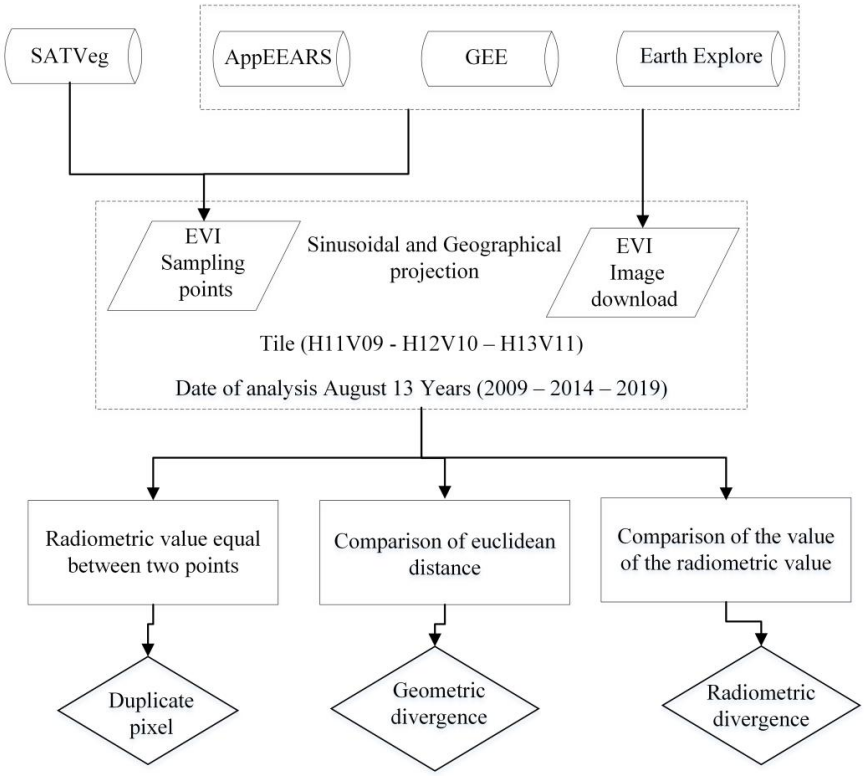

Figure 2. Flowchart of the steps for the analysis of radiometric and geometric divergences of virtual remote sensing platforms.

3 RESULTS AND DISCUSSION

3.1 Geometric pixel positioning

By comparing the pixel position of the images in the geographical projection, it became evident a displacement of the grid between the platforms (Figure 3). With respect to the sinusoidal projection, there was no apparent geometric displacement of pixels, as well as pixel size with the value of 231.322 m for all images evaluated.

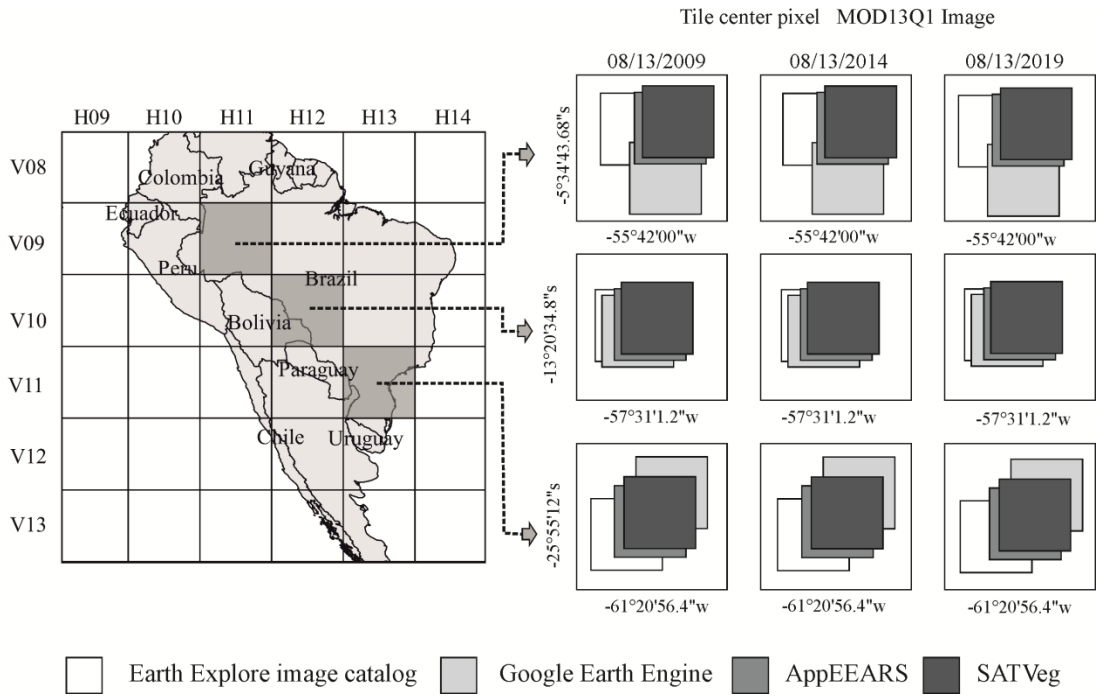

Figure 3. Pixel positioning in relation to the initial point of analysis for SATVeg, GEE, AppEEARS and the Earth Explore image catalog for MOD13Q1 images exported from virtual platforms in the geographic projection.

The largest shift occurred in the GEE images, in geographical projection, relative to the positioning of the Earth Explore MOD13Q1 image (Table 2). MODIS images of the GEE are clipped into blocks, maintaining the resolution and native projection of the image (Gorelick et al., 2017). At the moment the image is exported, the output projection takes place by sinusoidal pattern by making use of the resampling of the nearest neighbor, being the pixel size defined by the user. The resampling, in this case, was carried out in two occasions – one when making the blocks for storage, in the database, and the other one when the file was clipped, reprojected and exported. This entire process may have had an impact on this displacement, since, according to Seong

(2003), image processing on global or continental scales, as well as the projection and resampling procedures may bring significant distortions of pixels.

Table 2. Euclidean distance from the pixel center of the MOD13Q1 image from Earth Explore in relation to the pixel centers for the SATVeg, GEE and AppEEARS virtual platforms in different MODIS scenarios

		Distance (m)		
Projection	Tile	GEE	SATVeg	AppEEARS
Geographic	H11V09	108	58	60
	H12V10	27	62	71
	H13V11	98	129	134
Sinusoidal	H11V09	0	-	0
	H12V10	0	-	0
	H13V11	0	-	0

For AppEEARS, the projection and clipping of the image was also performed by making use of the resampling of the nearest neighbor; however, the pixel size was defined on the platform based on the central pixel of the native image, and this same size was adopted in the projection of the other images (Neeley, 2018). In the bibliographical research of this paper, no information was obtained about the resampling of the projection of images for SATVeg; however, due to the proximity in distance to AppEEARS, the same resampling method may have been used for the projection of images.

3.2 Points with data duplicity

Duplicate information was observed at two points, in a row, for all platforms when the geographical projection was used. In the MOD13Q1 base image from Earth Explore, there has been resampling for geographical projection, originating this same effect, which was also found. This duplication may be occurring because of the geographic projection, as found in the images, which preserves the angles; thus the areas of pixels will be larger as the latitude values increase (Downs, 2016). The conversion of metric values into degrees changes the angulation of the scenario, and in

order to readjust the pixels, the method of the nearest neighbor was used – if two sampled pixels are close to the same pixel, there will be a duplication of information.

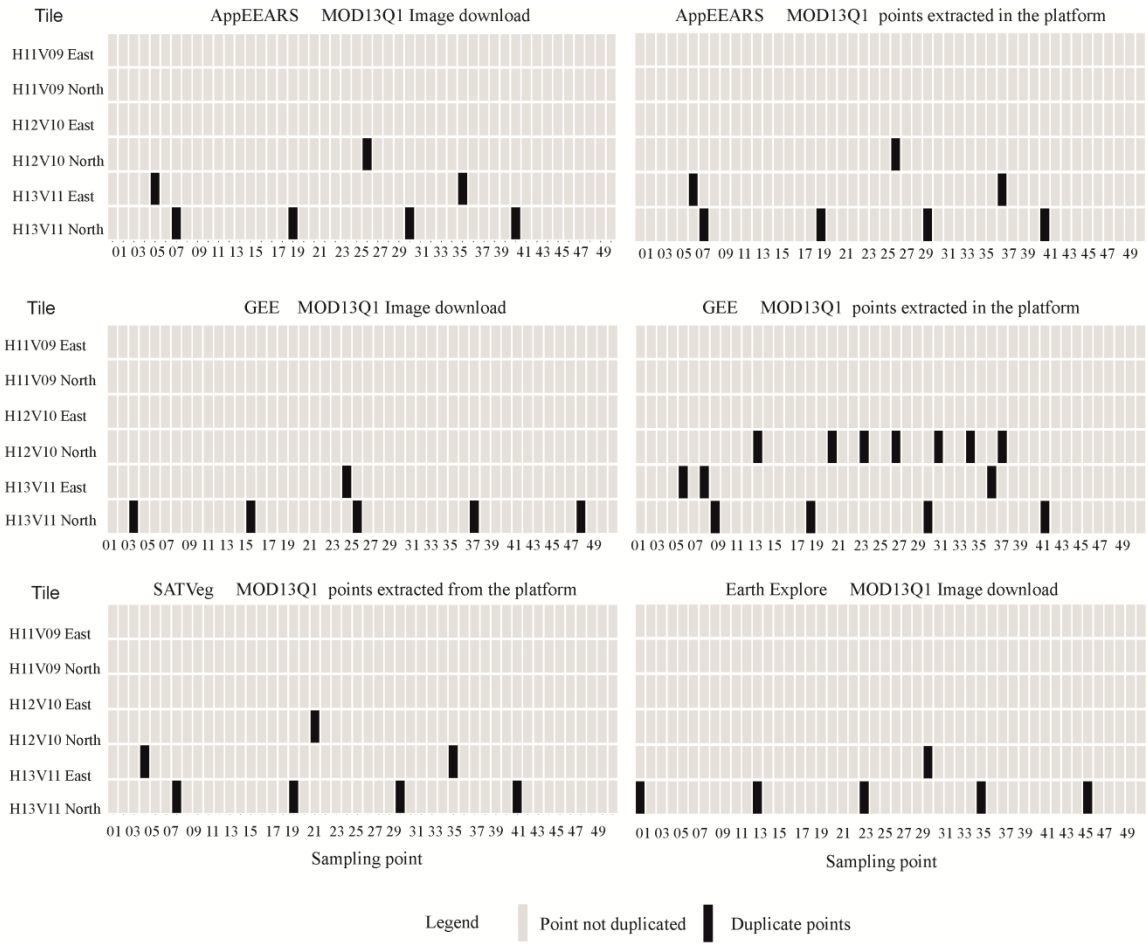

Figure 4. Points with duplicate pixel values in more than one date (north and east) for the MOD13Q1 EVI on SATVeg, GEE, AppEEARS and the Earth Explore image catalogs MOD13Q1 images exported from virtual platforms in the geographic projection.

The problems in pixel duplication are linked to the increased level of data processing that can accumulate the errors involved in each process, such as image projection. Depending on the projection of images, there may be duplicate or even omitted pixel values. These effects can be minimized by adopting an equivalent projection (Seong et al., 2002), Seong (2003), suggested strict care in the projection of global and continental scale image data, as model errors are sensitive to the effects of the curvature of the Earth, which leads to a significant loss in precision.

For the images that were downloaded on the GEE and AppEEARS platforms, there was the projection of images, and this was a factor for pixel duplicity, but this

information is not clear about the values collected by points, directly on the platforms. However, the analyses of the points in AppEEARS, in the duplicate locations, are in the same order as the duplicate points in the downloaded image, which suggests that the MOD13Q1 images, in this platform, are in the geographical projection, with the resampling carried out by the method of the nearest neighbor.

The GEE allows the visualization of the MOD13Q1 images in its own interface, as well as the SATVeg, which allows the visualization of the pixel in which the information is extracted. Note the difference in geometry between the platforms – the SATVeg pixel is in the form of a well-defined square, whereas the GEE the pixel refers to a distorted square (Figure 5).

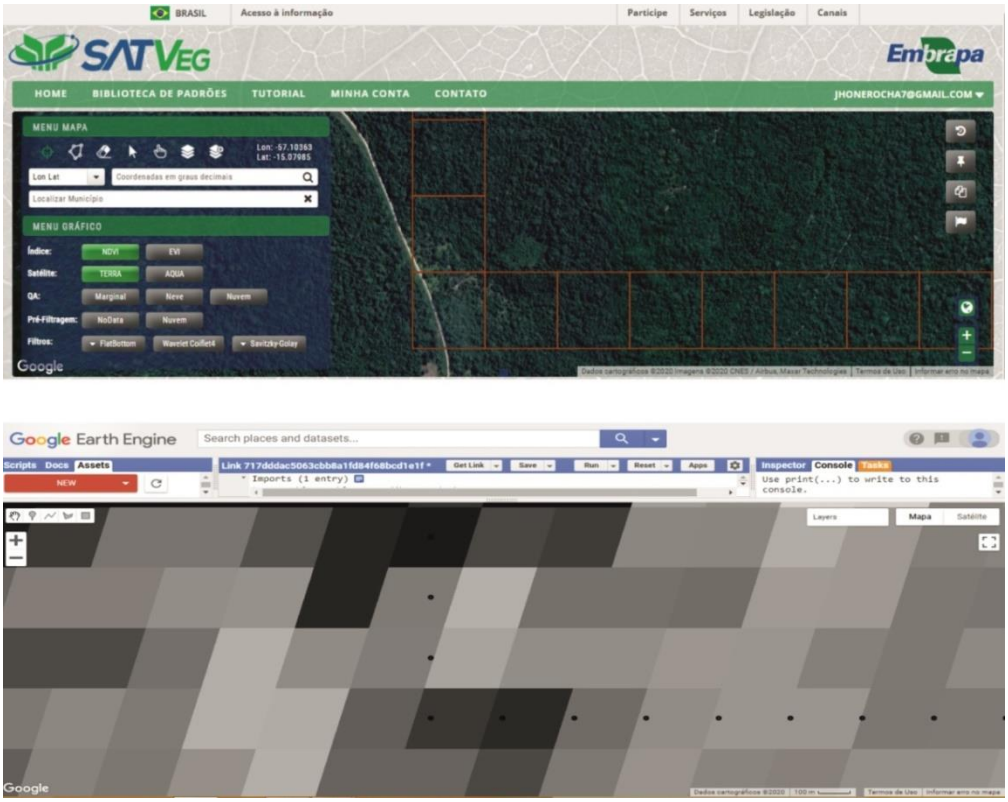

Figure 5. SATVeg platform interface (sample pixels) and Google Earth Engine EVI image of the MOD13Q1 product.

As specified in SATVeg, the pixels arranged in its interface were projected in geographic coordinates, being the pixels in quadratic form. The divergence in GEE geometry may be attributed due to the fact that image presented is in sinusoidal projection, and the project coordinates in geographical projection. When adopting a standard projection for a geographical database project and in the case that the image is

not in the same projection, an adaptation of the image occurs, causing distortion of the pixel in the image.

Regarding the evaluation of the platforms by making use of the sinusoidal projection, no duplicate pixel values were found. In this projection, there was no need for image resampling, thus maintaining the integrity of the radiometric values of the images. Sinusoidal projection presented less image distortion and was the most indicated for MODIS products (Cammalleri and Vogt, 2015; Roy et al., 2016). Mulcahy (1998) examined several projections in the world map, among them the geographical and the sinusoidal ones, being the latter the only projection that did not duplicate pixels.

3.3 Pixel displacement

Radiometric divergences were found in all platforms that made use of geographical projection. For the points collected in the east direction, more coincident values were found, but more divergent values were also found. Regarding the values collected in the northern direction, a large part of the values was found in the condition of displaced values. In other words, for the same geographic coordinates, there were different radiometric values (Figure 6).

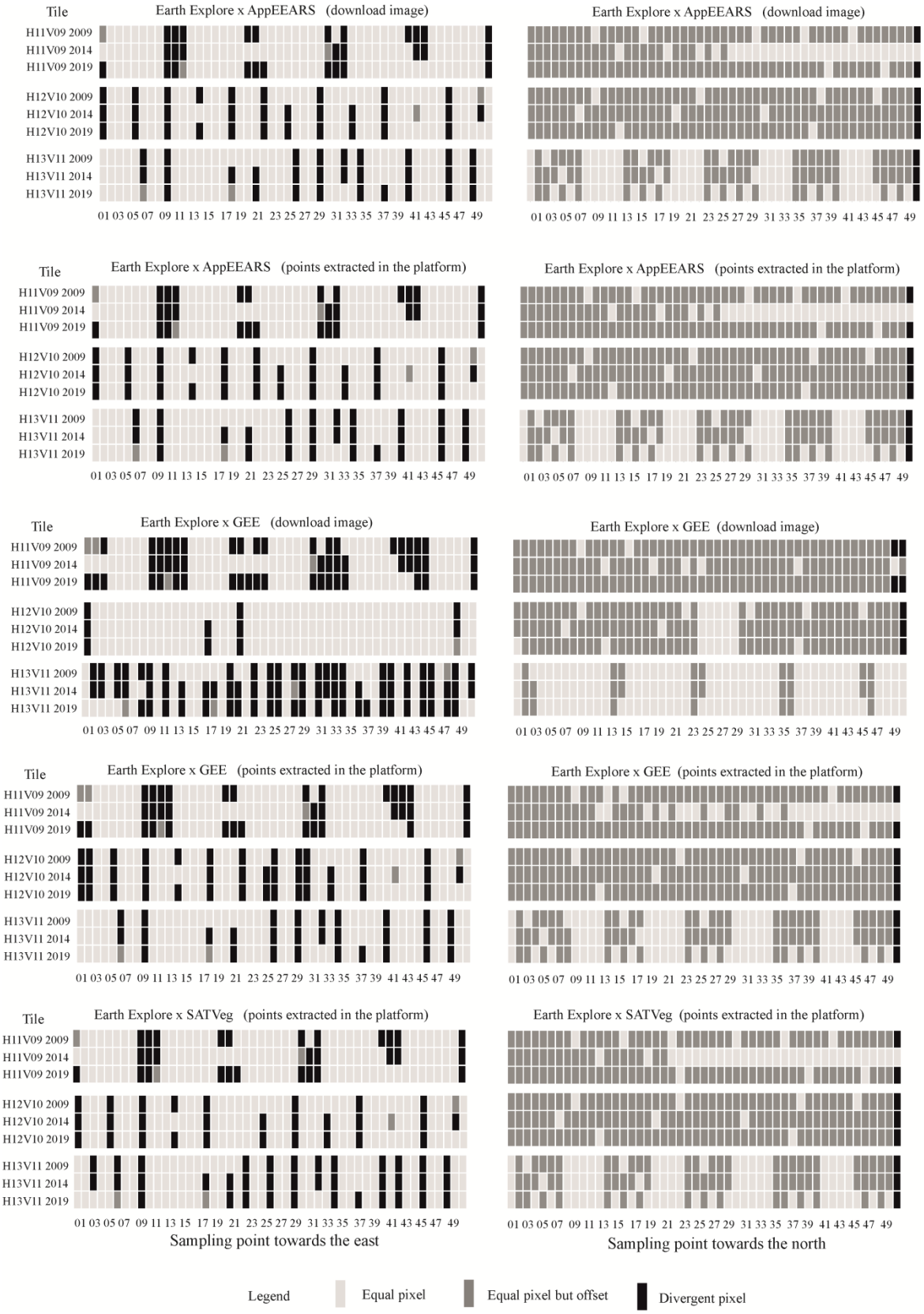

Figure 6. Divergences of the GEE, SATVeg and AppEEARS platforms in relation to the Earth Explore image catalog for the points in east and north directions for the years of 2009, 2014 and 2019, MOD13Q1 images exported from virtual platforms in the geographic projection.

The analysis of the duplicate pixels already indicated the divergences between the platforms due to the regions where this problem occurred, in different positions. Regarding the geometry of the platform images, there was also a difference between the distance from the center of the reference pixel. The results reinforce that the resampling of the images, with the closest neighbor method, can take into account divergent pixels between the platforms, because this would justify the divergent values.

By making use of the sinusoidal projection, there was no divergence between the radiometric values between the platforms (Table 3). The analysis of duplicity and geometry of the platforms also indicated that the radiometric values would tend to be equal. The image distortions in the sinusoidal projection are of less impact, but not without errors and are more likely to occur in high altitude regions (Li et al., 2018). Vannan et al. (2009) reported that in sinusoidal grid pixels the resolution may vary by up to 0.5% between grid lines.

Table 3. Total survey of the condition of pixels for the dates 08/12/2004, 08/13/2010 and 08/12/2016 for the SATVeg, GEE platforms in relation to Earth Explore.

Pixel condition	AppEEARS (download image)					
	Geographic Projection			Sinusoidal Projection		
	H11V09	H12V10	H13V11	H11V09	H12V10	H13V11
Equal pixel	51%	42%	69%	100%	100%	100%
Equal pixel, but offset	38%	47%	20%	0%	0%	0%
Divergent pixel	11%	11%	11%	0%	0%	0%
Pixel condition	AppEEARS (points extracted in the platform)					
	Geographic Projection			Sinusoidal Projection		
	H11V09	H12V10	H13V11	H11V09	H12V10	H13V11
Equal pixel	51%	42%	69%	-	-	-
Equal pixel, but offset	38%	47%	20%	-	-	-
Divergent pixel	11%	11%	11%	-	-	-
Pixel condition	GEE (download image)					
	Geographic Projection			Sinusoidal Projection		
	H11V09	H12V10	H13V11	H11V09	H12V10	H13V11
Equal pixel	67%	87%	64%	100%	100%	100%
Equal pixel, but offset	14%	8%	10%	0%	0%	0%
Divergent pixel	19%	4%	26%	0%	0%	0%
Pixel condition	GEE (points extracted in the platform)					

	Geographic Projection			Sinusoidal Projection		
	H11V09	H12V10	H13V11	H11V09	H12V10	H13V11
Equal pixel	50%	39%	67%	-	-	-
Equal pixel, but offset	38%	46%	23%	-	-	-
Divergent pixel	12%	14%	10%	-	-	-
Pixel condition	SATVeg (points extracted in the platform)					
	Geographic Projection			Sinusoidal Projection		
	H11V09	H12V10	H13V11	H11V09	H12V10	H13V11
Coincident	56%	44%	64%	-	-	-
Displaced	35%	46%	23%	-	-	-
Missing	9%	10%	13%	-	-	-

3.4 Features of Virtual Remote Sensing Platforms

The options available in the sensing platforms addressed offer a number of advantages and disadvantages, and it is up to the user to define the one that best suits him or her in relation to general aspects of diagnosed problems (Table 4).

Table 4. Specification of Virtual Remote Sensing Platforms as to their structure and diagnosed problems

	AppEEARS	GEE	SATVeg
Programming Knowledge	No	Java / Python	No
Possibility of time series filters on the images	No	Yes	Yes
Image selection based on information quality	Yes	Yes	Yes
Download multiple points in a single file		Yes	No
Download multi-platform images	Yes	Yes	No
Time to update new data	Daily	Daily	15 days
Processing Metadata	No	No	No
Image Metadata	Yes	Yes	No
Processing time for requested information	Until 24 hrs	Until 5 min	Immediate
Duplicate pixel	* Yes	* Yes	Yes
Displaced Pixel	* Yes	* Yes	Yes
Absent pixel	* Yes	* Yes	Yes

*Using geographic projection in relation to Earth Explore images

AppEEARS, within the platforms evaluated, was the reference in information because it was under the domain of the native data holder – the U.S. government agency – the National Aeronautics and Space Administration (NASA). Data collection in this platform proved to be advantageous because it was possible to evaluate the descriptive

statistics of the data, obtaining media values, quartiles, and departure of the EVI values, in addition to pixel quality. On the other hand, when the time series was extracted from all points together, the process was too prolonged, reaching a 24-hour period to make the results available.

GEE stands out for its wide range of satellite images, which can be accessed through fast filtering capabilities. These tools allow a selection of images based on spatial, temporal and quality of information acquired Gorelick et al.(2017), besides enabling the rapid interactive visualization of results throughout the development of the algorithm. In one case when the algorithm is fixed, it can be portioned out and implemented in applications that grant direct access to the GEE database.

However, these tools are data structure for potential applications, that is, an experienced programmer is expected to implement these algorithms so that he or she can reach novice users. According to Gorelick et al.(2017), there is a limitation of data processing in this platform, with the maximum processing duration of 270 seconds and 40 simultaneous requests. Each user can store 250 gigabytes of files in the platform. In case of exceedance of this limit, there is an option to hire the Google Cloud service to host them. It is recommended for GEE to perform all the desired processing on the platform and to export only the results, thus tends to decrease the geometric displacement caused by the resampling of the images. The exportation of the point values alone showed little difference from the data from AppEEARS.

In the SATVeg, the great advantage over the other platforms was that it presented information with a high level of processing, which can facilitate the recognition of vegetation and climatic property standards in relation to Interpret. Nevertheless, these differences between platforms are frequently induced by the composition of the database, with SATVeg consisting of vector files that are weaker in nature and of lower operating cost than raster files.

In these registration databases, the images should be processed, which, in turn, will be resampled, which may result in geometric displacement at this time. In this case, a vector mesh should serve as the basis for extracting the raster values of images, and these meshes may contain alignment errors, with images and impact on the divergences demonstrated between the platforms.

The AppEEARS platform counts on the government resource from the partnership of the National Aeronautics and Space Administration (NASA) and the

United States Geological Survey (USGS) (Maiersperger, 2017). Google Domain GEE contains someone's company's physical and financial structure to maintain a platform. However, Brazil has been undergoing political and economic changes and reducing investment in teaching and research, forcing funding agencies to try alternative forms of economic management to keep the platform dynamic and update. Maintaining a platform with cadastral databases tends to be less costly for storing data processing compared to a database provided with the images.

3.5 Final considerations

Virtual access platforms are technological dissemination initiatives that permit the engagement of the most diverse users, which can provide improvements in research and practical applications with regard to orbital remote sensing. However, it is necessary to make the appropriate metadata for the interpreter to make proper use of these tools. The reliability of information is critical for these platforms to be used as decision-making tools. With regard to alert monitoring, an error in the data acquired can cause logistical, personal and fiscal losses and affects the credibility of the sector coming.

The use of remote sensing of virtual platforms for multi-areas of knowledge can be applied to contribute to the understanding of phenomena that are rarely targets of attention to Geoinformation professionals. But for the insertion of these users, it is necessary to democratize these platforms, similar to Series View and SATVeg, which require little knowledge in the area and still provides a spectral pattern platform; however, it does not offer an implementation of analytics applications, such as GEE beyond the absence of metadata.

The availability of metadata can help integrate information from virtual platforms, given the fact that in possession of internal procedures it is possible to adapt its configuration of data acquisition and thus reduce its divergences. For example, the definition of an analysis location, the spectral behavior pattern in SATVeg, the reliability of values in AppEEARS, and data from other platforms in GEE. All platforms have advantages and disadvantages, and it is necessary to evaluate the reliability of available information science, given that platform with divergent data can result in distinct interpretations.

3.6 Future work

Other remote sensing web platforms could be evaluated in future work. As example, Goddard Earth Sciences Data and Information Services Center (GES DISC) has developed Goddard, an online interactive visualization infrastructure and analytics tool (Giovanni) that allows access to MODIS data in a similar time series as AppEEARS (Berrick et al., 2008). Vegetation Spot is the platform that provides vegetation indexes from Système Probatoire of the Observation of Earth (SPOT5), with land surface monitoring since 2002, being the advantage of this platform its daily data feed, but in return its spatial resolution is of 1 km (Maisongrande et al., 2004). These programs can also be measured in future studies by making use of the standard to compare database information right away from crude images and validate radiometric and geometric values presented.

4 CONCLUSION

It was observed that there were divergences in raster values among the online remote sensing platforms. The resampling method in the reprojection of the images may be causing these divergences because this problem is only diagnosed in the geographical projection. By proving the divergences between the virtual platforms, it can be demonstrated that there is a need for a reference to validate the information. The metadata of the processing of virtual platforms are relevant information to ensure that the reproduction of searches in different sources offer the same result.

COMPUTER CODE AVAILABILITY

Name of code: Virtual Platform Analysis

- Developers: Jonathan da Rocha Miranda
- Contact details – Department of Agricultural Engineering, Federal University of Lavras, University Campus, PO Box 3037, ZIP Code: 37200-000, Lavras, Minas Gerais, Brazil, email: jhonerocha@estudante.ufla.br
- Year first available: 2020

- Hardware required: Virtual Platform Analysis was run on a computer with 4 cores (2.4 GHz each) and 4 GB
- Software required: Virtual Platform Analysis was interpreted with Spyder and needs pandas, NumPy, seaborn and matplotlib packages
- Program language: the code is written in Python 3.6
- Program size: 153 kb
- Details on how to access the source code: the source files of the Virtual Platform Analysis can be downloaded from GitHub: https://github.com/jonathanrocha71/virtual_platform_analysis.git

ACKNOWLEDGEMENTS

The authors wish to thank (i) the Department of Agricultural Engineering of the Federal University of Lavras (UFLA) for providing office space and infrastructure to achieve this article, as well as (ii) the Foundation for Supporting Research of the State of Minas Gerais (FAPEMIG).

REFERENCES

Antunes, J.F.G., Lamparelli, R.A.C., Rodrigues, L.H.A., 2015. Assessing of the sugarcane cultivation dynamics in São Paulo state by MODIS data temporal profiles. Eng. Agrícola 35, 1127–1136.

Bailey, J.E., Chen, A., 2011. The role of Virtual Globes in geoscience. Comput. Geosci. 37, 1–2. https://doi.org/10.1016/j.cageo.2010.06.001

Berrick, S.W., Leptoukh, G., Farley, J.D., Rui, H., 2008. Giovanni: a web service workflow-based data visualization and analysis system. IEEE Trans. Geosci. Remote Sens. 47, 106–113.

Buras, A., Rammig, A., Zang, C.S., 2019. Quantifying impacts of the drought 2018 on European ecosystems in comparison to 2003. arXiv Prepr. arXiv1906.08605.

Cammalleri, C., Vogt, J., 2015. On the role of land surface temperature as proxy of soil moisture status for drought monitoring in Europe. Remote Sens. 7, 16849–16864.

Chang Seong, J., 2003. Modelling the accuracy of image data reprojection. Int. J. Remote Sens. 24, 2309–2321. https://doi.org/10.1080/01431160210154038

Daubechies, I., 1990. The Wavelet Transform, Time-Frequency Localization and Signal Analysis. IEEE Trans. Inf. Theory 36, 961–1005. https://doi.org/10.1109/18.57199

Downs, R., 2016. Adventures in Academic Cartography: A Memoir, Imago Mundi. Bar Scale Press Syracuse, NY. https://doi.org/10.1080/03085694.2016.1107416

Duval, E., Hodgins, W., Sutton, S., Weibel, S.L., 2002. Metadata principles and practicalities. D-Lib Mag. 8, 1082–9873. https://doi.org/10.1045/april2002-weibel

Embrapa Informática Agropecuária, 2016. Sistema de análise temporal da vegetação: SATVeg. Campinas, [2016].

Gallo, K., Schumacher, P., Boustead, J., Ferguson, A., 2019. Validation of Satellite Observations of Storm Damage to Cropland with Digital Photographs. Weather Forecast. 34, 435–446.

Georgi, C., Spengler, D., Itzerott, S., Kleinschmit, B., 2018. Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. Precis. Agric. 19, 684–707. https://doi.org/10.1007/s11119-017-9549-y

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Remote Sensing of Environment Google Earth Engine : Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Gorry, P.A., 1990. General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method. Anal. Chem. 62, 570–573. https://doi.org/10.1021/ac00205a007

Justice, C.O., Vermote, E., Townshend, J.R.G., Defries, R., Roy, D.P., Hall, D.K., Salomonson, V. V., Privette, J.L., Riggs, G., Strahler, A., Lucht, W., Myneni, R.B., Knyazikhin, Y., Running, S.W., Nemani, R.R., Wan, Z., Huete, A.R., Van Leeuwen, W., Wolfe, R.E., Giglio, L., Muller, J.P., Lewis, P., Barnsley, M.J., 1998. The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. IEEE Trans. Geosci. Remote Sens. 36, 1228–1249. https://doi.org/10.1109/36.701075

Karami, M., Rangzan, K., Saberi, A., 2013. Using GIS servers and interactive maps in spectral data sharing and administration: Case study of Ahvaz Spectral Geodatabase Platform (ASGP). Comput. Geosci. 60, 23–33. https://doi.org/10.1016/j.cageo.2013.06.007

Kastens, J.H., Brown, J.C., Coutinho, A.C., Bishop, C.R., Esquerdo, J.C.D.M., 2017.

Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil. PLoS One 12, e0176168.

Li, J., Chen, S., Qin, W., Useya, J., Zhen, Z., Wang, Y., 2018. A Fast Reprojection Method for MODIS Products with Sinusoidal Projection. J. Indian Soc. Remote Sens. 46, 1563–1567. https://doi.org/10.1007/s12524-018-0794-y

Li, R., Shi, J., Ji, D., Zhao, T., Plermkamon, V., Moukomla, S., Kuntiyawichai, K., Kruasilp, J., 2019. Evaluation and Hydrological Application of TRMM and GPM Precipitation Products in a Tropical Monsoon Basin of Thailand. Water 11, 818.

Macedo, R. de C., Schmitt Filho, A.L.S., Farley, J.C., Fantini, A.C., Cazella, A.A., Sinisgalli, P.A. de A., 2018. Land use and land cover mapping in detailed scale: a case study in Santa Rosa de Lima-SC. Bol. Ciências Geodésicas 24, 217–234.

Maiersperger, T., 2017. AppEEARS: A Simple Tool that Eases Complex Data Integration and Visualization Challenges for Users, in: AGU Fall Meeting Abstracts.

Maisongrande, P., Duchemin, B., Dedieu, G., 2004. VEGETATION/SPOT: an operational mission for the Earth monitoring; presentation of new standard products. Int. J. Remote Sens. 25, 9–14.

Melo, M.R. da S., Rocha, J.V., Manabe, V.D., Lamparelli, R.A.C., 2018. Intensity of land use changes in a sugarcane expansion region, Brazil. J. Land Use Sci. 13, 182–197.

Miranda, L.S., Lipp-Nissinen, K.H., 2017. Delimitation of permanent preservation areas of Paurá Lagoon (Middle Coast of Rio Grande do Sul, Brazil) using multitemporal satellite image analysis. Rev. Gestão Costeira Integr. Integr. Coast. Zo. Manag. 17, 65–75.

Mulcahy, K.A., 1998. Spatial Data Sets and Map Projections: An Analysis of Distortion.

Münch, Z., Gibson, L., Palmer, A., 2019. Monitoring Effects of Land Cover Change on Biophysical Drivers in Rangelands Using Albedo. Land 8, 33.

Neeley, S., 2018. Analyzing Earth Data with NASA's AppEEARS Tool to Improve Research Efficiency, in: AGU Fall Meeting Abstracts.

Quenzer, R., Friesz, A.M., 2015. AppEEARS: Simple and Intuitive Access to Analysis Ready Data, in: AGU Fall Meeting Abstracts.

Roy, D.P., Li, J., Zhang, H.K., Yan, L., 2016. Best practices for the reprojection and

resampling of Sentinel-2 Multi Spectral Instrument Level 1C data. Remote Sens. Lett. 7, 1023–1032.

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., Ohno, H., 2005. A crop phenology detection method using time-series MODIS data. Remote Sens. Environ. 96, 366–374. https://doi.org/10.1016/j.rse.2005.03.008

Seong, J.C., Mulcahy, K.A., Usery, E.L., 2002. The sinusoidal projection: A new importance in relation to global image data. Prof. Geogr. 54, 218–225. https://doi.org/10.1111/0033-0124.00327

Su, Y., Slottow, J., Mozes, A., 2000. Distributing proprietary geographic data on the World Wide Web - UCLA GIS database and map server. Comput. Geosci. 26, 741–749. https://doi.org/10.1016/S0098-3004(99)00130-2

Trentin, A.B., Trentin, C.B., Saldanha, D.L., Kuplich, T.M., 2018. ANÁLISE DE SÉRIES TEMPORAIS MODIS E TRMM EM PLANTIOS DE EUCALIPTO. Mercator 17.

Vannan, S.K.S., Cook, R.B., Holladay, S.K., Olsen, L.M., Dadi, U., Wilson, B.E., 2009. A web-based subsetting service for regional scale MODIS land products. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2, 319–328.

Wardlow, B.D., Kastens, J.H., Egbert, S.L., 2006. Using USDA crop progress data for the evaluation of greenup onset date calculated from MODIS 250-meter data. Photogramm. Eng. Remote Sens. 72, 1225–1234.

Yildirim, F., Kaya, A., 2008. Selecting map projections in minimizing area distortions in GIS applications. Sensors 8, 7809–7817. https://doi.org/10.3390/s8127809

Zhu, L. feng, Wang, X. feng, Zhang, B., 2014. Modeling and visualizing borehole information on virtual globes using KML. Comput. Geosci. 62, 62–70. https://doi.org/10.1016/j.cageo.2013.09.016

Zurier, S., 1996. Geographic information system look around-you find GISes wherever feds are in Government. Comput. News 15, 61.

PAPER 3 - DETECTION OF COFFEE BERRY NECROSIS BY DIGITAL IMAGE PROCESSING OF LANDSAT 8 OLI SATELLITE IMAGERY

Journal standards: *International Journal of Applied Earth Observation and Geoinformation*, ISSN 0303-2434, https://doi.org/10.1016/j.jag.2019.101983.

Jonathan da Rocha Miranda[a], Marcelo de Carvalho Alves[a], Edson Ampélio Pozza[b], Helon Santos Neto[b]

[a] Department of Agricultural Engineering at the Federal University of Lavras, University Campus, P.O.Box 3037, 37200-000, Lavras, Minas Gerais, Brazil

[b] Plant Pathology Department, Federal University of Lavras, University Campus, P.O.Box 3037, 37200-000, Lavras, Minas Gerais, Brazil

ABSTRACT: Coffee berry necrosis is a fungal disease that, at a high level, significantly affects coffee productivity. With the advent of surface mapping satellites, it was possible to obtain information about the spectral signature of the crop on a time scale pertinent to the monitoring and detection of plant phenological changes. The objective of this paper was to define the best machine learning algorithm that is able to classify the incidence CBN as a function of Landsat 8 OLI images in different atmospheric correction methods. Landsat 8 OLI images were acquired at the dates closest to sampling anthracnose field data at three times corresponding to grain filling period and were submitted to atmospheric corrections by DOS, ATCOR, and 6SV methods. The images classified by the algorithms of machine learning, Random Forest, Multilayer Perceptron and Naive Bayes were tested 30 times in random sampling. Given the overall accuracy of each test, the algorithms were evaluated using the Friedman and Nemenyi tests to identify the statistical difference in the treatments. The obtained results indicated that the overall accuracy and the balanced accuracy index were on an average around 0.55 and 0.45, respectively, for the Naive Bayes and Multilayer Perceptron algorithms in the ATCOR atmospheric correction. According to the Friedman and Nemenyi tests, both algorithms were defined as the best classifiers. These results demonstrate that Landsat 8 OLI images were able to identify an incidence of the coffee berry necrosis by means of machine learning techniques, a fact that cannot be observed by the Pearson correlation.

Keywords: Data mining; spectral behavior; accuracy; *Colletotrichum* ssp; atmospheric correction.

1 INTRODUCTION

Coffee farming has always played a prominent role in Brazilian commodities. In the 2017/2018 harvesting, Brazilian coffee production accounted for about 32.4% of the world market for *in natura* coffee (IOC, 2018). The technologies employed from

planting to commercialization are being increasingly demanded mainly with the advent of precision agriculture, which can provide in gaining productivity.

Within the mechanisms adopted in crop management, knowledge of tools that monitor pests and diseases is essential. Regarding coffee diseases, the Coffee Berry Necrosis – CBN – is one of the main coffee diseases, since it has a direct action on coffee productivity. CBN, in most cases, is related to fungi of the genus *Colletotrichum*, which has reported a reduction of up to 80% in productivity (Griffiths et al., 1971; Varzea et al., 2002).

Although the fungus acts on fruits, the presence of *Colletotrichum* spp. in coffee branches cause changes in the normal stem and leaf structure due to physiological disorders that are associated with disease onset. Among them, the most common are high pending fruit loads, nutritional deficiency, physical and chemical impediments in the soil (Paradela Filho, 2001). Sera et al., (2005), observed a negative correlation between the increased incidence of *Colletotrichum* ssp. and the vegetative vigor, which was evaluated in the visual perception of the plant, observing the leaf tone and branch dryness.

The incidence of the disease can change the density of the canopy and the leaf area, factors that can be identified by spectral signature mainly in the infrared region (Franke and Menz, 2007). The combination of different wavelengths may be able to detect diseases by multispectral sensors, given that the disease signals may influence peculiarly the spectral signature of the target (Mahlein et al., 2013). Therefore, multispectral satellite imagery can aid the detection and control of the pathogen due to its ability to infer in physiological aspects of plants (Lopresti et al., 2015).

The use of orbital images for detection, quantification and classification of coffee diseases has been used and improved over time (Chemura et al., 2018a, 2017; Price et al., 1993; Tucker et al., 2013). Accurate and reliable detection of diseases is facilitated by highly sophisticated and innovative methods of data analysis that lead to new insights derived from sensor data for complex plant-pathogen systems (Mahlein, 2016).

Machine learning algorithms make no assumptions about frequency distribution and are becoming increasingly popular to classify remote sensing data, which rarely have normal distributions (Belgiu and Drăgu, 2016). It is estimated that the techniques of machine learning can find a classifier capable of identifying the incidence of coffee

berry necrosis based on Landsat 8 OLI images, once the relation of the spectral signature of the coffee canopy under the effect of coffee berry necrosis incidence is known.

This paper aimed to define which methods of atmospheric correction combined with machine learning techniques can approximate the process of evaluating the disease in the field data.

2 MATERIAL AND METHODS

The analyzed area is located in the southern region of Minas Gerais, in a coffee crop, in the municipality of Carmo do Rio Claro, centered at coordinates of latitude 21°00'28" South of Ecuador and longitude 46°01'30" West of Greenwich (Fig 1). The planting of coffee (Coffea arabica L.) cultivar Acaiá 474/19 was arranged with a spacing of 3.6 m between rows and 0.70 m between plants in a total area of 11 hectares. The crop was irrigated through drip irrigation with management based on the water demand, measured through properly installed tensiometer batteries.

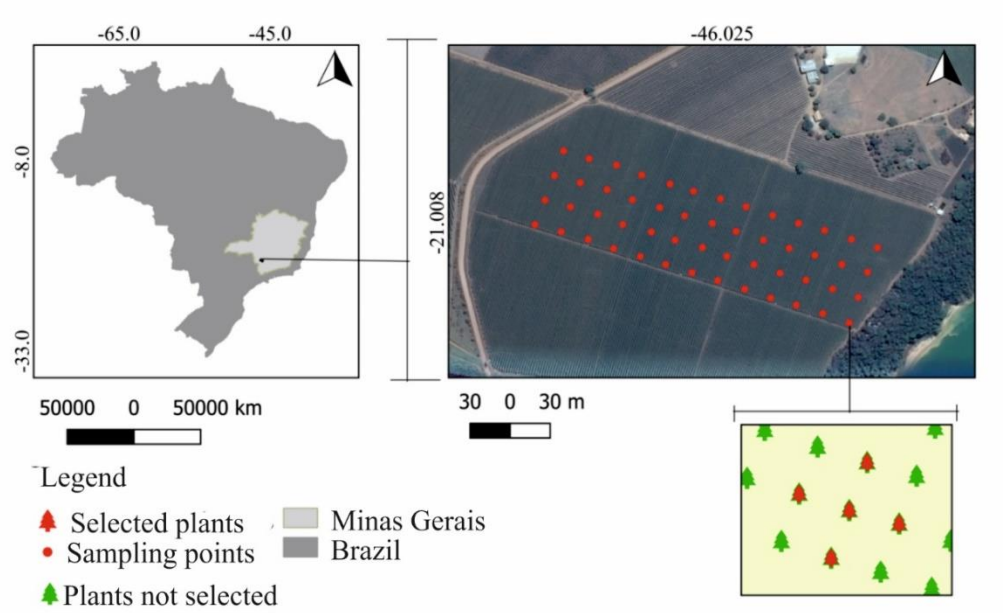

Fig 1. The spatial location of the analyzed area highlighted the distribution of the sample set, emphasizing the five coffee plants selected in each sample.

2.1 Berry necrosis assessment in the field

The sample mesh was composed of georeferenced points in a spacing of 40 to 40 meters measured in the field with a GPS TRIMBLE 4600 LS ® and Total Station Leica TC600 ® (Fig 1). In each georeferenced point five plants were assessed, and in each plant two branches in the third middle of the plant canopy were randomly defined and marked using a wire in order to assess the same branch in all evaluated periods. The percentage of reproductive nodes with diseased berries in relation to reproductive nodes of the branch was evaluated (Santos Neto, 2017).

During disease assessment, necrotic berry samples were collected and sent to the diagnostic and control laboratory of plant diseases of the Federal University of Lavras (UFLA), where it was possible to observe the presence of the fungus *Colletotrichum* ssp. The isolates were morphologically characterized as belonging to *Colletotrichum gloeosporioides* species complex and had its pathogenicity proved on coffee berries.

The assessment of the berries disease was carried out in three periods (December, 15, 2013; January, 18 and February, 26, 2014), corresponding to the period of grain development according to the scale of evaluation of phenological stages proposed by Pezzopane et al., (2003). All assessments were considered over the months in a single sample, which allowed to separate the intensity of coffee berry necrosis in 4 bands of classes defined by the quartile (0 - 25, 25 - 50, 50 - 75, 75 - 100% of the total number of samples), which corresponded to 32 samples per established class.

2.2 Constitution of the sample mesh to remote sensing analysis

The sample mesh was composed of georeferenced points in a spacing of 40 by 40 meters collected by means of a GPS TRIMBLE 4600 LS ® and Total Station Leica TC600 ®. In these points, it is considered a buffer of 7.2 meters of radius corresponding to the spacing between two lines of planting.

At these points, its framing was not considered in the pixels of the Landsat 8 OLI image. In this sense, a criterion of selection of the points was put into practice, in which it guarantees that it is representative of a single pixel. The criterion of point selection was established in the condition that the polygon of the buffer contained in a single pixel of the Landsat 8 OLI image, discarding all points in an intersection between

two pixels. The resampling was carried out by moving the pointer to the closest position to the center of the pixel in which it contained to reorganize a mesh structure coincident with the images (Fig 2).

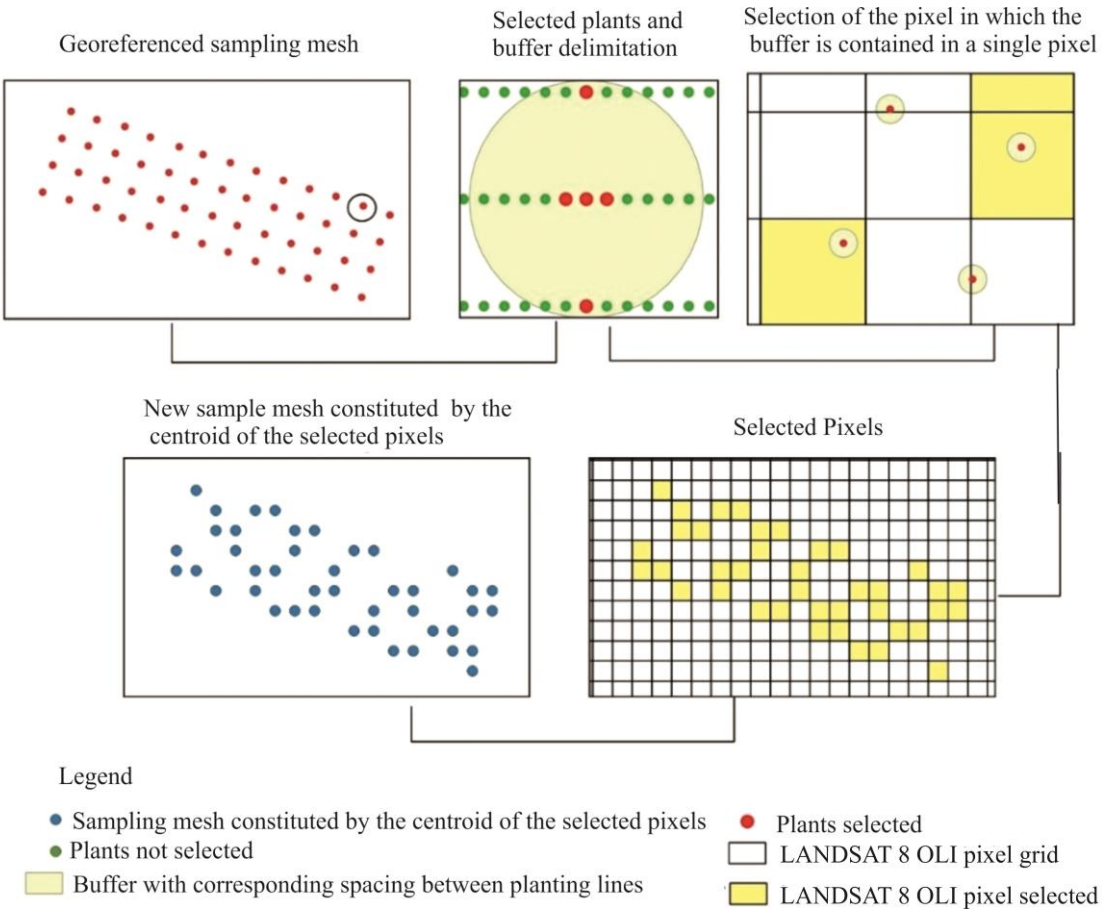

Fig 2. Scheme of the sample selection process coincident with Landsat 8 OLI images.

2.3 Atmospheric corrections

Images from the Landsat 8 OLI satellite at orbit 219 points 75 which are freely available through the Earth Explore online interface. These images were obtained as close as possible to the date of the field sampling to minimize the phenological transformations that occur over time, which may prevent the direct correlation between the intensity of the necrotic berries and its spectral characteristics. The selected images were collected from December, 6, 2013, January, 23 and February, 24, 2014 (Table 1 and Table 2).

Table 1. Images Landsat 8 OLI used in the process and metadata information

Date	12/06/2013	01/23/2014	02/24/2014
Land cloud cover (%)	17.38	23.39	38.89
Sun Azimuth angle	76.08	94.12	100.16

Table 2. Description of Landsat 8 OLI used products

Bands	Wavelength (μm)	Description
B2	0.452-0.512	Blue
B3	0.533-0.590	Green
B4	0.636-0.673	Red
B5	0.851-0.879	Near Infrared
B6	1.566-1.651	Short-wave infrared 1
B7	2.107-2.294	Short-wave infrared 2

Three atmospheric correction models were used to compare atmospheric correction methods more suitable to characterize the features in the coffee canopy. It has been used the Dark Object Subtraction (DOS), Atmospheric and Topographic Correction for Satellite Imagery (ATCOR) and Second Simulation at the Satellite Signal in the Solar Vector Spectrum (6SV).

The atmospheric correction by the DOS method (Chavez, 1988) performed considering the histogram of the image of the region of smaller wavelength, in which for Landsat 8 OLI images refers to the band of blue of a wavelength of 0.43 μm. The input information were the raw images and their metadata. Based on this information, the atmospheric interference in each spectral band was estimated followed by calculations for the transformation of the digital number into radiance values and then for reflectance values. All procedures of the equations can be consulted in the study of Chavez (1988).

For ATCOR atmospheric correction, the algorithm proposed by Richter (1996) consisted in the entry of a fog-free image, cloud shadow and full pixel mask. However, in the absence of this information, the azimuth and zenith angle information were used as the basis, the calibration coefficients contained in the metadata Image. Consequently, the top atmosphere reflectance (TOA) and the cloud mask were defined.

To mitigate the effects of cloudiness that masked the actual reflectance, the values of the highest and lowest brightness pixels and the maximum magnitude (in pixels) present in each cloud have been adjusted. The altered adjustment values

followed the criterion in which the current mask overlies the maximum on the cloud cover.

The adjustment of the scene lighting conditions was established based on the digital elevation model (DEM) imagery from the Shuttle Radar Topographic Mission (SRTM). This feature allows the radar, transmittance and irradiance values to be obtained in conjunction with the TOA image. This process was performed iteratively to recover the surface reflectance value for each pixel.

Considering the effects of aerosols on the atmosphere, the rural model was selected to represent the aerosol conditions that are not influenced by urban or industrial sources. It is a product of the reactions between atmospheric gases and the effects of the dust particles (Richter and Schläpfer, 2011, 2003).

For atmospheric correction 6SV (Vermote et al., 2016), the product already processed by NASA has been used. Atmospheric data for the latest 6SV models use the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor as the source of information for applying the atmospheric correction to Landsat 8 OLI images.

With these reflectance values corrected by atmospheric effects, it was possible to perform radiometric transformations in order to enhance the vegetation information. The normalized difference vegetation index (NDVI) proposed by Rouse (1973), the improved vegetation index (EVI) proposed by Huete et al. (1997) and the normalized difference water index (NDWI) proposed by Gao (1996) have been used. (Equations 1, 2 and 3).

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \quad (1),$$

$$EVI = G * \frac{\rho_{nir} - \rho_{red}}{L + \rho_{nir} + C_1 * \rho_{red} - C_2 * \rho_{blue}} \quad (2),$$

$$NDWI = \frac{\rho_{nir} - \rho_{swir1}}{\rho_{nir} + \rho_{swir1}} \quad (3),$$

where: ρ_{nir} is the band reflectance of Near-Infrared; ρ_{red} is the reflectance band of red; ρ_{swir1} is the band reflectance of short-wave infrared 1; ρ_{blue} is the band reflectance of blue; L is the soil adjustment factor adopted for value 1; C_1 is the coefficient for the aerosol effect adopted in value 6; C_2 also refers to aerosol, however the adopted value was 7.5; G refers to gain factor of 2.5.

2.4 Machine Learning process

We have used the Python-based learning algorithm Naive Bayes (John and Langley, 1995), Random Forest (Breiman, 2001) and the Multilayer Perceptron (Hinton, 1990).

For the Naive Bayes algorithm, the classification that has been performed started by estimating the probabilities of each class, followed by the calculation of the respective mean, so the algorithm constructed the covariance matrices forming the discriminant function for each type according to Bhargavi and Jyothi (2009).

In the classification by Random Forest, we used a set of 500 decision trees that were formed by values of the set of data sampled by bootstrap. According to Lawrence et al. (2006), estimation errors tend to stabilize before this reached tree number. The entropy criterion was used in tree hierarchization to reduce the randomness of the classifier.

However, the classifier by the Multilayer Perceptron configured so that the found errors were less than 0.000001 in the classification, or that reached a maximum iteration of 1000, the optimization of the weights adjusted the errors based on the Adam stochastic function gradient (Kingma and Ba, 2014).

The algorithms were validated by the 10-fold stratified cross-validation method, which consisted of the iteration number of the algorithms in which each round is a new set of training data, and test was changed and organized with the same amount of repetitions of each sample class. The results of the classification were defined by the final mean of alliterations, as recommended by the authors (Hall et al., 2009).

2.5 Evaluation of machine learning models

We performed 30 tests of the algorithms, modifying the sample set by randomizing the generating seed from 1 to 30. In each analysis, the global accuracy, user accuracy, producer accuracy and Balanced Accuracy (BAC) have been calculated. The global accuracy was defined by the number of the correctness of the error matrix by the total of evaluated samples. The accuracy of the user is associated with the commission error, in which the committed error is attributed to a pixel that does not belong to the true class. The producer's accuracy is associated with the omission error,

which occurs when we fail to map a pixel in the true class and the BAC is the average specificity and sensitivity.

The best classifier defined by the method of Friedman and Nemenyi was the one that performs a non-parametric analysis of variance for a single factor of variation and makes comparisons between independent samples by ordering the data by increasing values; and then the original values are replaced by the order number in a set of ordered series.

The global accuracy values of each test ranked in increasing order among the classifiers in different atmospheric corrections. If there are no statistically significant differences between the two classifiers, they will be connected in the diagram by a straight line (Rodríguez et al., 2010). All discussed processing is expressed in the flowchart below (Fig 3).

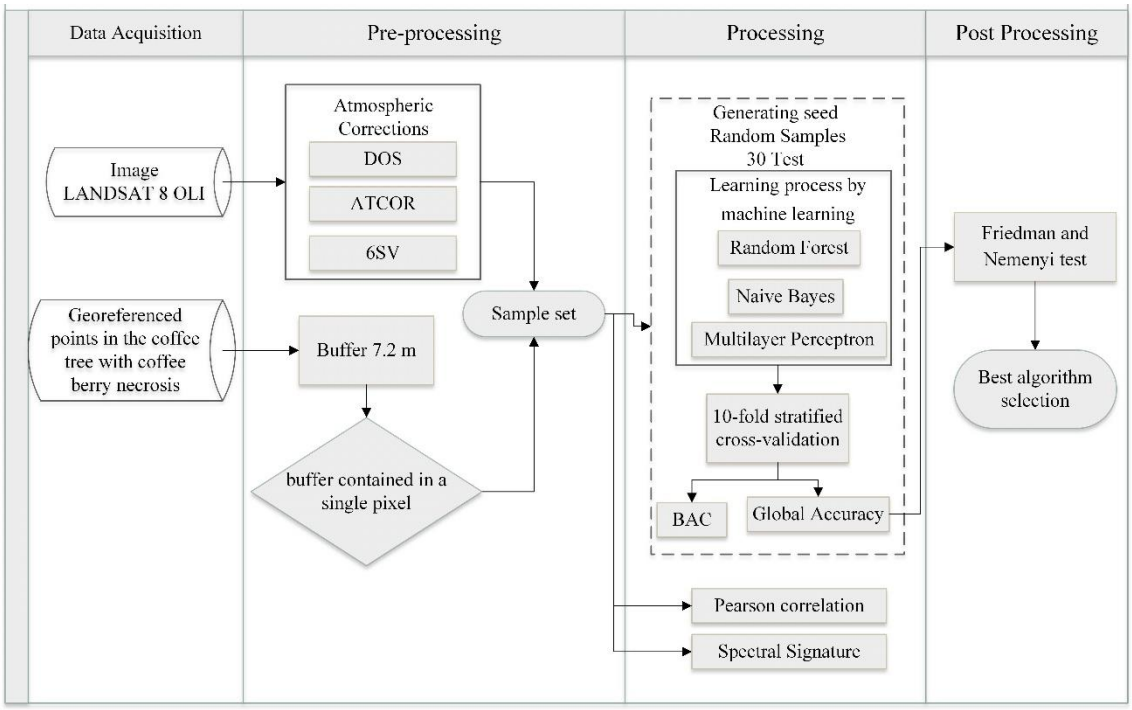

Fig 3. Scheme of the methodology used to choose the best combination of data for the machine learning process in the classification of the incidence of coffee berry necrosis.

3 RESULTS AND DISCUSSION

On-site monitoring showed a peak infestation in February, 2014 (Fig 4 and 5). During this period, berries were in the phenological phase of expansion, in which there

was a greater proliferation of the attack of the pathogen on the reproductive nodes. We believe that the occurrence of diseases in the coffee crop could be influenced by factors related to the pathogen's virulence, as well as host resistance, climatic conditions and crop management (Maia et al., 2013).

There were isolated points in February in which the incidence was lower than that found in January. This was a normal diseased because berries tend to fall. The infestation of the *Colletotrichum gloeosporioides* in an advanced stage presents necrotic centers that when reaches the leaves and berries of the coffee tree tend to cause the early fall beside the dry of the branches. However, in these situations, the same value of incidence of the previous evaluation was considered, since the berry fall did not represent if there was a decrease in the disease (Fernandes and Vieira Junior, 2015; Paradela Filho, 2001).

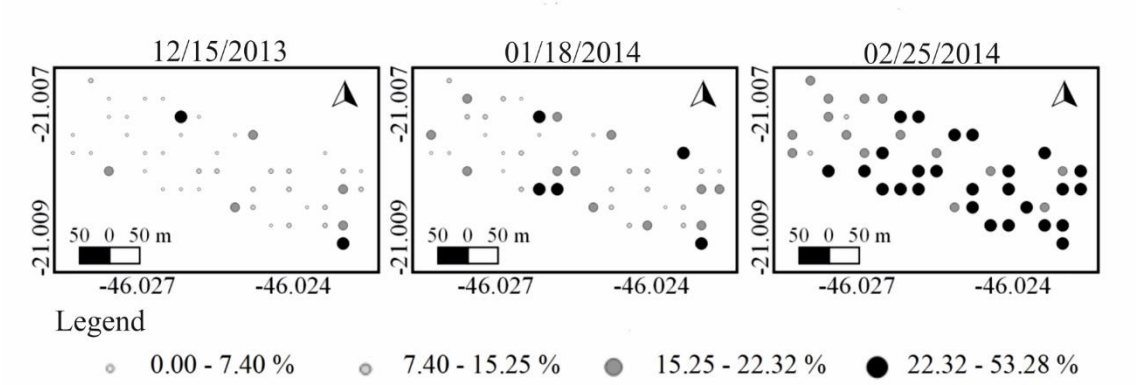

Fig 4. Quantile spatial distribution of the incidence of coffee berry necrosis field data throughout December (2013), January and February (2014).

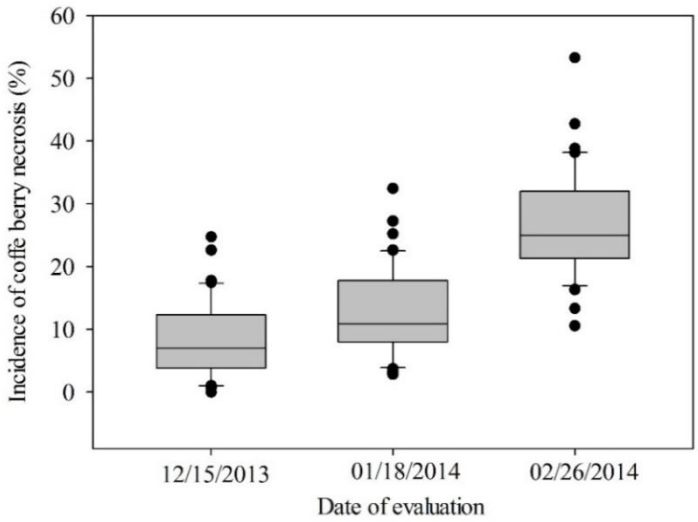

Fig 5. Box plot of field data evaluations of the coffee berry necrosis incidence during December (2013), January, and February (2014).

In the evaluation of the atmospheric correction models, there was not a discrepant difference in the visualizations in true color compositions (RGB-432), with an exception of December (Fig 6). According to the quality of the pixel that is available in the images corrected by 6SV, specifically in this month, there was a higher concentration of aerosol and the presence of clouds. These factors mainly affect the bands in the spectrum in the region of the visible that corresponds to the bands of 0.43, 0.56 and 0.69 μm. The effects of the aerosol concentration on bands centered at shorter wavelengths (0.43 μm) make the surface reflectance generally small, with a robust aerosol signal. In this case, there will be a greater Rayleigh-type dispersion and gas absorption of electromagnetic energy (Vermote et al., 2016). The other configurations of colored compositions presented similar tonalities and visual appearance among themselves. It was the color-accurate color composition that performed the best contrast between the atmospheric corrections.

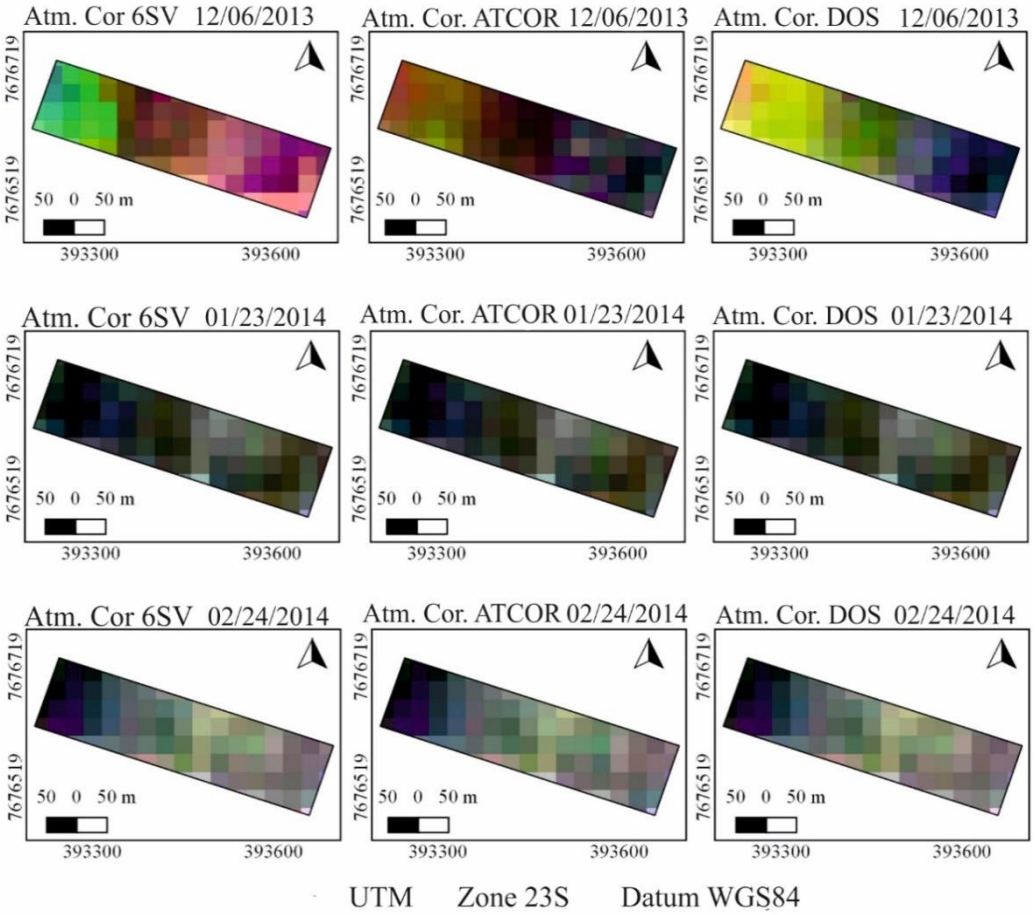

Fig 6. RGB-321 Landsat 8 OLI color composition for the 6SV, ATCOR and DOS atmospheric correction methods at the dates closest to the field data assessments of coffee berry necrosis incidence.

The reflectance in the visible region was lower in February than in previous months, and this aspect can be observed independently regarding the atmospheric correction method and coffee berry necrosis incidence (Fig 7). Because the crop is relatively young, approximately three-years old, the plant is in full development, being the month of February with greater leafing, which consequently absorbed more energy in the visible region.

The atmospheric correction in the DOS method presented higher values of reflectance compared to the other methods, possibly because the algorithm does not consider meteorological factors in the calculation. The DOS, the atmospheric effects scaled by the distribution of the histogram of the assigned image in the bands of shorter wavelengths and formulates a linear equation of atmospheric correction for each band (King, 2003). This method was not sufficient to mitigate the influence of the atmosphere on the image.

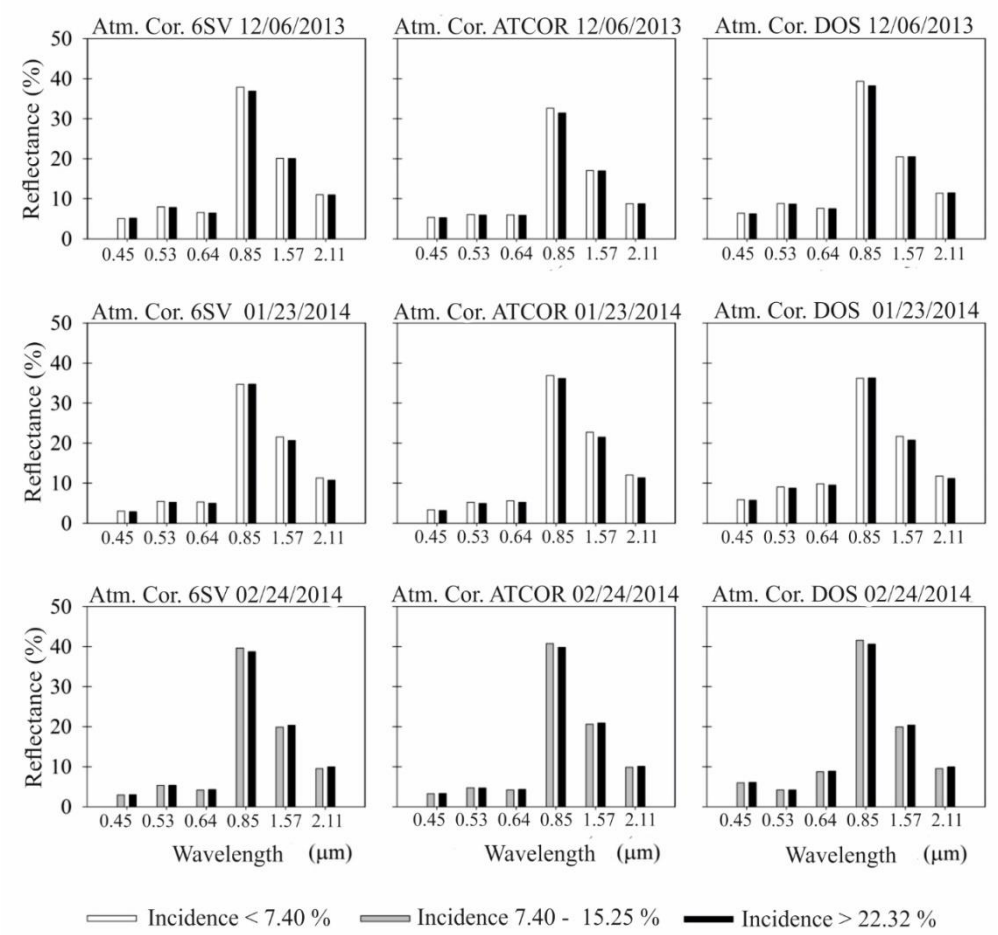

Fig 7. Spectral signature of the upper and lower classes of the coffee berry necrosis incidence during December, January and February for the 6SV, ATCOR, and DOS atmospheric correction methods.

According to Martinelli et al. (2015), pathogens cause a reduction in chlorophyll content in leaves due to necrotic lesions, tending to change in spectral signature, causing a change in the value of index vegetation. It has been observed that small differences between changes in reflectance values (Fig 7) and vegetation indices between incidence less than 7.40% and greater than 22.32% did not exceed 10%, but even subtle there was a signal that fruit necrosis decreases in the value of vegetation indices (Table 3).

Table 3. Vegetation index values for the lowest and highest coffee berry necrosis incidence classes in the analyzed period in images of different atmospheric corrections.

Atmospheric correction	Index vegetation	12/06/2013 Incidence (%)		01/23/2014 Incidence (%)		02/24/2014 Incidence (%)	
		<7.40	>22.32	<15.40	>22.32	<7.40	>22.32
6SV	NDVI	0.716	0.685	0.741	0.738	0.816	0.796
	NDWI	0.325	0.278	0.249	0.235	0.342	0.302
	EVI	0.581	0.532	0.515	0.510	0.618	0.603
ATCOR	NDVI	0.742	0.700	0.579	0.575	0.864	0.842
	NDWI	0.332	0.281	0.267	0.253	0.361	0.321
	EVI	0.600	0.531	0.439	0.434	0.819	0.795
DOS	NDVI	0.701	0.662	0.737	0.734	0.819	0.800
	NDWI	0.331	0.273	0.246	0.232	0.338	0.299
	EVI	0.535	0.474	0.517	0.512	0.626	0.612

The relationship between coffee berry necrosis and reflectance had higher correlations in the medium infrared region and also in the index vegetation, specifically for December and February (Fig 8). In this region, the internal scattering of the electromagnetic radiation occurred as a consequence of its interaction with the leaf mesophyll, external factors such as the disease can be altering the water and air relation in the mesophyll, resulting in a lower reflectance (Franke and Menz, 2007).

Similar results were found by Boechat et al. (2014), who evaluated the spectral signature of the bean white mold using a spectroradiometer field data. The authors observed that in the near-infrared region, the leaves infected by the fungi presented lower reflectance than in healthy leaves, caused by the destruction of plant tissues during colonization of the leaves, as correlations between NDVI and a white mold severity, which were not statistically understood between loading and grain maturation.

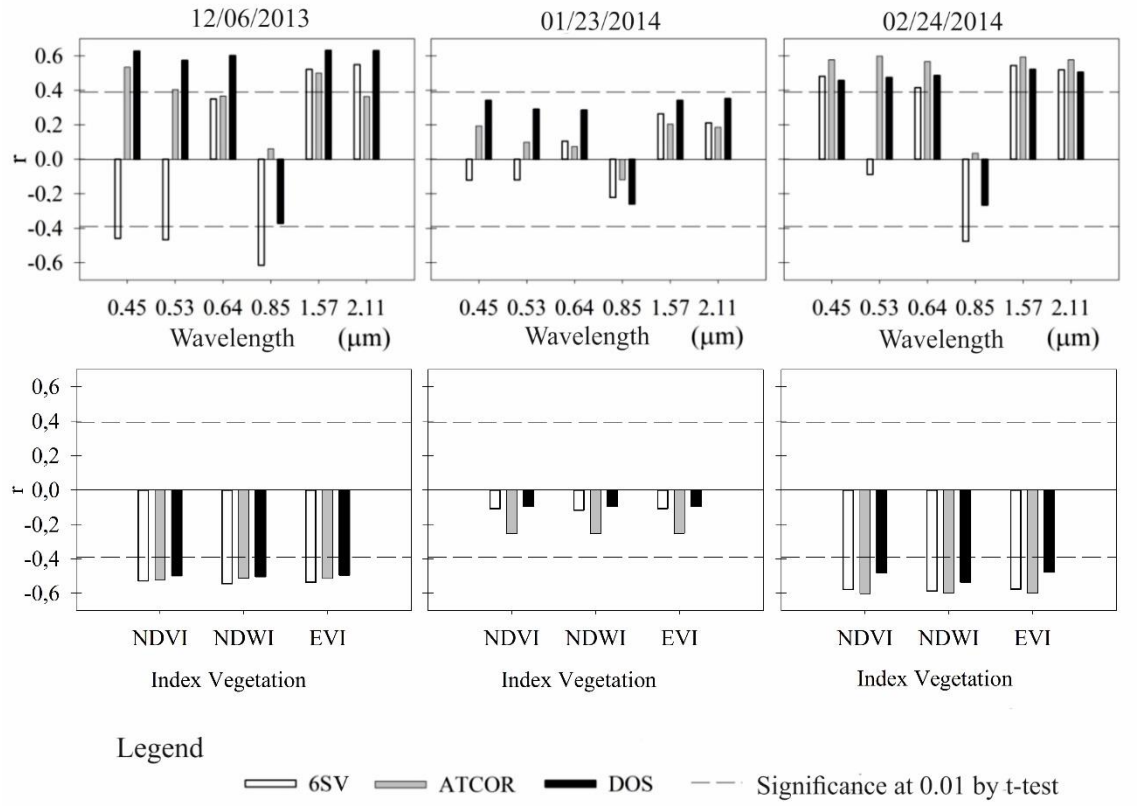

Fig 8. Pearson correlation coefficient (r) for the reflectance of the wavelengths and index vegetation of the Landsat 8 OLI images between the incidence of coffee berry necrosis ($\alpha \leq 0.01$).

Other studies also report the existence of the relationship between reflectance and pathogen infection in plants. In a study by Zhao et al. (2014), the authors evaluated the effect of the severity of yellow rust (*Puccinia striiformis*) on wheat by hyperspectral reflectance of a spectrophotometer. The correlations were highest in the visible region and reached the maximum correlation at about 0.85 with the disease. In the near and medium infrared region, there was a decrease in the mean correlation to 0.45. Prabhakar et al. (2011), in a study on the cotton stress attacked by green spittlebugs (*Hemiptera cicadellidae*), obtained a high negative correlation in the near-infrared region with the pest, reaching a value of -0.77, which corroborated the average R^2 determination index of 0.68 in the NDVI ratio and the leafhopper. In the infrared medium and in the visible region, the correlation was high and positive, with the maximum in the red wavelength reaching approximately 0.79. These studies revealed the potential of using spectral signatures as indicative of physiological disturbances in crops, from satellite imagery, such as Landsat 8 OLI.

Concerning January, the correlations were not significant in the test. In this case, the hypothesis is that the reflectance relations and the incidence of the coffee berry necrosis were not linear. In some instances, not even hyperspectral satellites operating in narrow and specific bands obtained linear relationships, probably due to the spectral mixing components (Chemura et al., 2018b).

From the use of the machine learning algorithms, it was possible to identify the level of the coffee berry necrosis incidence as a function of the reflectance of the Landsat 8 OLI bands for every month in the analysis, even in different atmospheric corrections, surpassing the answers obtained by the Pearson correlation. Specifically for biological analysis, modern machine learning techniques were capable of describing patterns that exceed the estimates determined by conventional statistical methods, such as regression and linear correlation (Ma et al., 2014).

In the results concerning the identification of the class intensity of the coffee berry necrosis as a function of the reflectance, the ATCOR atmospheric correction presented the higher values of accuracy and BAC index was compared to the 6SV and DOS ones (Tab 4). The inclusion of the SRTM images in the corrections by the ATCOR allowed a reference of illumination originated by the terrain effect, which guaranteed standardization between the images of different evaluated dates. According to Richter and Schläpfer (2011), this mode of atmospheric correction is especially important in cases of multi-temporal, multi-sensor or multi-condition images and must be standardized so that they can be compared.

Table 4. Global Accuracy and the Balanced Accuracy (BAC) mean index of the 30 random seed evaluations of the Multilayer Perceptron, Random Forest and Naive Bayes algorithms in detecting the incidence of coffee berry necrosis.

	Atm. Cor. 6SV		Atm. Cor. ATCOR		Atm. Cor. DOS	
All Bands						
Machine Learning Algorithm	Global Accuracy	BAC	Global Accuracy	BAC	Global Accuracy	BAC
Multilayer Perceptron	0.516	0.523	0.577	0.589	0.509	0.518
Random Forest	0.498	0.504	0.487	0.494	0.499	0.503
Naive Bayes	0.549	0.539	0.585	0.577	0.463	0.459
All Bands and Index vegetation						
Multilayer Perceptron	0.520	0.521	0.579	0.580	0.505	0.505
Random Forest	0.504	0.507	0.493	0.493	0.519	0.521
Naive Bayes	0.534	0.526	0.545	0.540	0.475	0.463

Vermote et al. (2016) specified that the 6SV correction was limited to uniform and flat targets. Tan et al. (2013) indicated that in cases of rough terrain, the topographic effects may introduce interference in the reflectance promoted by the shading of the image that in this method are not counted in the 6SV atmospheric correction. The application of the DOS method in soft or smooth undulating areas may lead to the underestimation of the atmospheric effects on the images due to the low presence of shading (Chavez, 1988).

The global accuracy and BAC using all bands were similar with respect to analysis of all bands with the addition of vegetation indices. All indices have in common the use of the near-infrared band that obtained significant correlation only with the application of 6SV atmospheric correction, so its relationship with fruit necrosis was not enough to increase the accuracy of machine learning models. By the principle of parsimony, which refers to choosing a smaller set of information to elucidate the problem according to Powers and Turk (2012), it has been decided to use only the spectral bands for further analysis.

Considering that the disease level control has an incidence of 5%, being tolerable up to 12%, the primary attention in the accuracy was given to 0 to 7.4% incidence class. The Multilayer Perceptron algorithm had the highest efficiency classification in images corrected by the 6SV method, and the accuracy was around 0.6. Regarding the user accuracy, the most significant was 0.76, determined by the Naive Bayes in the atmospheric correction 6SV. This adjustment can be advantageous, once the sites that have not been affected yet by the pathogen were identified in the images, which makes targeting the regions possible, in which they need an application of fungicides.

In the lowest incidence class interval between 7.40 and 15%, the highest producer accuracy and user accuracy (of 0.8 and 0.43) was found by the Multilayer Perceptron algorithm in the ATCOR atmospheric correction, respectively (Table 5). These results may aid field data monitoring, in which the indication of the intensity of the disease can be evaluated punctually, thus providing a rapid identification and a previous mapping of the places where a control measure should be taken. Decisions on the timely management of diseases in coffee are particularly important because they are closely linked to the yield losses (Martinelli et al., 2015).

Table 5. Producer and user accuracy for the coffee berry necrosis incidence classes obtained by the Multilayer Perceptron, Random Forest and Naive Bayes classifier algorithms based on the Landsat 8 OLI image reflectance in the 6SV, ATCOR and DOS atmospheric correction methods.

Machine Learning Algorithm	Incidence Class (%)	Atm. Cor. 6SV		Atm. Cor. ATCOR		Atm. Cor. DOS	
		Producer Accuracy	User Accuracy	Producer Accuracy	User Accuracy	Producer Accuracy	User Accuracy
Multilayer Perceptron	00.00 - 07.40	0.606	0.606	0.576	0.704	0.576	0.633
	07.40 - 15.25	0.452	0.333	0.806	0.439	0.452	0.318
	15.25 - 22.32	0.250	0.615	0.250	0.727	0.125	0.571
	22.32 - 53.28	0.727	0.585	0.758	0.735	0.758	0.521
Random Forest	00.00 - 07.40	0.545	0.486	0.576	0.514	0.576	0.500
	07.40 - 15.25	0.355	0.324	0.290	0.281	0.323	0.323
	15.25 - 22.32	0.281	0.333	0.344	0.393	0.375	0.414
	22.32 - 53.28	0.636	0.677	0.636	0.656	0.636	0.677
Naive Bayes	00.00 - 07.40	0.576	0.760	0.576	0.655	0.576	0.633
	07.40 - 15.25	0.613	0.396	0.742	0.404	0.355	0.324
	15.25 - 22.32	0.219	0.636	0.250	0.727	0.219	1.000
	22.32 - 53.28	0.758	0.556	0.758	0.781	0.818	0.466

The Friedman test indicated that there was a significant difference (p-value < 0.01) between the classifications, given that the Naive Bayes in the ATCOR correction suggested by the Nemenyi test is the best classifier (Fig 9). Although simple, Naive Bayes can often overcome some sophisticated classification methods (Farid et al., 2012). This algorithm based on the common assumption that all the characteristics are independent from one another, there is a tendency to be less tolerable to the changes in values in the spectral bands, being able to identify patterns of behavior even with few samples (Xu, 2018). This algorithm is known to be more sensitive to changes in the training set because it has a fixed structure and a small number of parameters (Rodríguez et al., 2013).

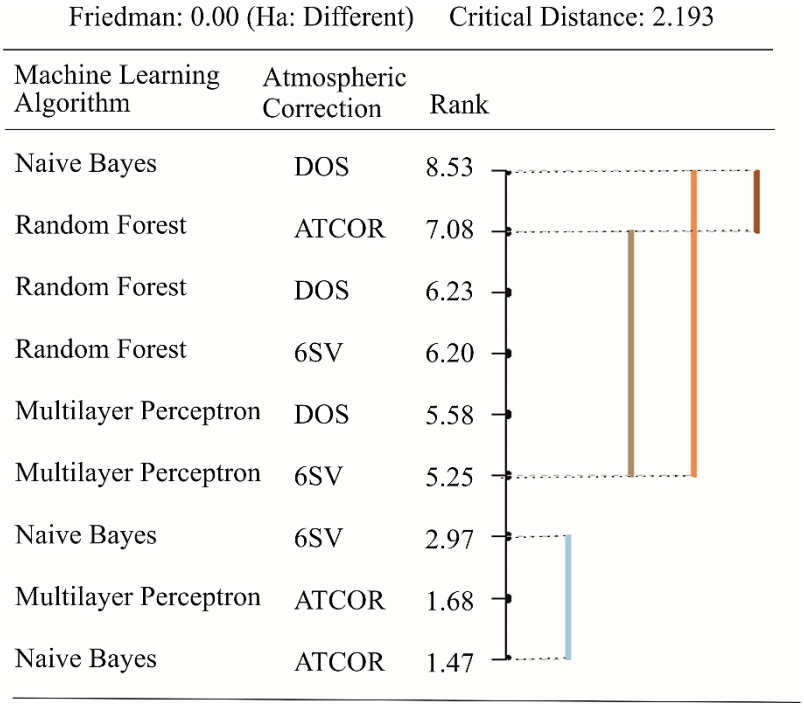

Fig 9. Friedman and Nemenyi test for the Multilayer Perceptron, Random Forest and Naive Bayes classifier algorithms to identify the coffee berry necrosis incidence classes based on the reflectance of Landsat 8 OLI images in the ATCOR, DOS, and 6SV atmospheric corrections.

The Multilayer Perceptron in the ATCOR correction also stood out among the best classifiers and statically equal to the Naive Bayes, as already pointed out the averages of the accuracy and BAC (Table 1) and the spatial distribution of the classes (Fig 10). Researches report on the efficiencies of classification techniques with the use of neural networks to produce results according to the presence/absence of the disease in crops and possibly severity levels. West et al. (2004), employing spectral images aboard a spectrograph mounted at the level of the spray bar, obtained increased performance of the classifier in 99% accuracy to differentiate healthy wheat from disease wheat using the algorithm of neural network perceptron multi-layered. Abdulridha et al. (2016) selected the appropriate wavelengths to correctly classify healthy trees of stressed trees with an accuracy of 98% through the neural network. Li et al. (2009) had an accuracy of 95% of the classification by neural networks for unhealthy rice stressed by rice diseases and healthy rice pests based on spectral leaf behavior.

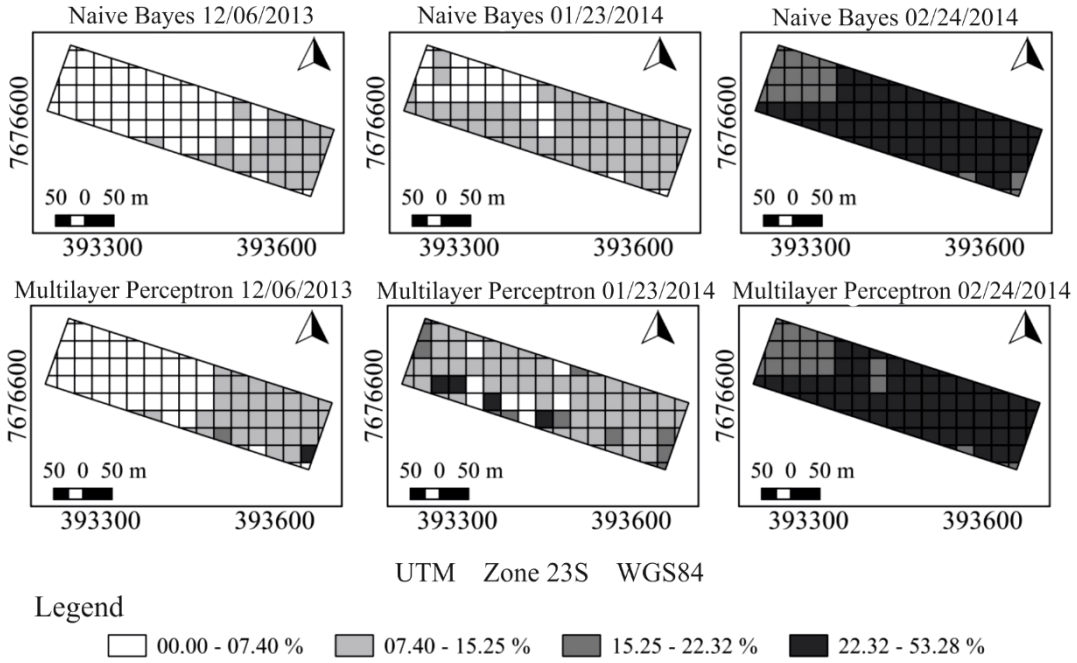

Fig 10. Incidence classification of coffee berry necrosis from the Naive Bayes and Multilayer perceptron algorithm in Landsat 8 OLI images used in the ATCOR atmospheric correction during December, January and February.

However, the performance of the classifiers depends on the arrangement of the data sets, being that each algorithm can have a better performance in each case. In this paper, Naive Bayes presented the best results, but this does not mean that it is the best classifier. As an example, Ma et al. (2015) identified the activities of relationships of biological structures in medicines by making use of Random Forest, obtaining the best performance in most of the data sets when compared to the classifier of deep learning by neural networks. In the study carried out by Russo et al. (2018), to predict compounds for endocrine-disrupting abilities, such as estrogen receptor binding, Random Forest once again had the best performance when compared to Naive Bayes and Multilayer Perceptron. Other researches evidence the potential of the Multilayer Perceptron. Were et al. (2015), mapping soil organic carbon variations, the Neural Networks algorithm obtained a superior performance of up to 36% when compared to Random Forest. This research evidenced the need for tests to choose the best learning algorithm of the machine adjusted to the data set.

4 CONCLUSION

The global accuracy and BAC were generally less than 0.60 for the trained data set. With a more robust database of samples from other coffee crops, it is believed that this result may be more accurate. The results are indicative that Landsat 8 OLI images may provide pertinent information for decision-making in agricultural planning, as well as for the timely application of pesticides. Machine learning tools were more efficient than Pearson's correlation to detect the incidence of coffee necrosis. The best classifier performance was Naive Bayes and Multilayer Perceptron in atmospheric images corrected by the ATCOR method.

ACKNOWLEDGEMENTS

The authors wish to thank (i) the Department of Agricultural Engineering of the Federal University of Lavras (UFLA) for providing office space and infrastructure to achieve this article, as well as (ii) the Foundation for Supporting Research of the State of Minas Gerais (FAPEMIG).

5 REFERENCES

Abdulridha, J., Ehsani, R., de Castro, A., 2016. Detection and Differentiation between Laurel Wilt Disease, Phytophthora Disease, and Salinity Damage Using a Hyperspectral Sensing Technique. Agriculture 6, 56. https://doi.org/10.3390/agriculture6040056

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bhargavi, P., Jyothi, S., 2009. Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. IJCSNS Int. J. Comput. Sci. Netw. Secur. 9, 117–122.

Boechat, L.T., de Carvalho Pinto, F. de A., de Paula, T.J., Queiroz, D.M., Teixeira, H., 2014. Detection of white mold in dry beans using spectral characteristics. Rev. Ceres 61, 907–915. https://doi.org/10.1590/0034-737X201461060004

Breiman, L., 2001. RANDOM FORESTS Leo. Mach. Learn. 45, 1–33. https://doi.org/https://doi.org/10.1023/A:1010933404324

Chavez, P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sens. Environ. 24, 459–479.

https://doi.org/10.1016/0034-4257(88)90019-3

Chemura, A., Mutanga, O., Dube, T., 2017. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. Precis. Agric. 18, 859–881. https://doi.org/10.1007/s11119-016-9495-0

Chemura, A., Mutanga, O., Odindi, J., Kutywayo, D., 2018a. Mapping spatial variability of foliar nitrogen in coffee (Coffea arabica L.) plantations with multispectral Sentinel-2 MSI data. ISPRS J. Photogramm. Remote Sens. 138, 1–11. https://doi.org/10.1016/j.isprsjprs.2018.02.004

Chemura, A., Mutanga, O., Sibanda, M., Chidoko, P., 2018b. Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. Trop. Plant Pathol. 43, 117–127. https://doi.org/10.1007/s40858-017-0187-8

Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M.A.A., Strachan, R., Mofizur, C., Hossain, M.A.A., Strachan, R., 2012. Expert Systems with Applications Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Syst. Appl. 41, 1937–1946. https://doi.org/10.1016/j.eswa.2013.08.089

Fernandes, C. de F., Vieira Junior, J.R., 2015. Coffee diseases, in: Embrapa Rondônia-Chapter of Scientific Book (ALICE). In: MARCOLAN, AL; ESPINDULA, MC (Ed.). Coffee in the Amazônia. Brasília, DF ….

Franke, J., Menz, G., 2007. Multi-temporal wheat disease detection by multi-spectral remote sensing. Precis. Agric. 8, 161–172. https://doi.org/10.1007/s11119-007-9036-y

Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. https://doi.org/https://doi.org/10.1016/S0034-4257(96)00067-3

Griffiths, E., Gibbs, J.N., Waller, J.M., 1971. Control of coffee berry disease. Ann. Appl. Biol. 67, 45–74. https://doi.org/10.1111/j.1744-7348.1971.tb02907.x

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl. 11, 10–18. https://doi.org/10.1145/1656274.1656278

Hinton, G.E., 1990. CONNECTIONIST LEARNING PROCEDURES. Mach. Learn. 555–610. https://doi.org/10.1016/B978-0-08-051055-2.50029-8

Huete, A.R., Liu, H.Q., Batchily, K. V, Van Leeuwen, W., 1997. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. Remote Sens. Environ. 59, 440–451. https://doi.org/https://doi.org/10.1016/S0034-4257(96)00112-5

John, G.H., Langley, P., 1995. Estimating Continuous Distributions in Bayesian Classifiers George, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence., UAI'95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 338–345. https://doi.org/10.1109/TGRS.2004.834800

King, R.B., 2003. Remote Sensing Geology., The Photogrammetric Record. Springer. https://doi.org/10.1046/j.0031-868x.2003.024_04.x

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization.

Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). Remote Sens. Environ. 100, 356–362. https://doi.org/10.1016/j.rse.2005.10.014

Li, B., Liu, Z., Huang, J., Zhang, L., Zhou, W., Shi, J., 2009. Hyperspectral identification of rice diseases and pests based on principal component analysis and probabilistic neural network. Nongye Gongcheng Xuebao/Transactions Chinese Soc. Agric. Eng. 25, 143–147. https://doi.org/10.3969/j.issn.1002-6819.2009.09.026

Lopresti, M.F., Di Bella, C.M., Degioanni, A.J., 2015. Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. Inf. Process. Agric. 2, 73–84. https://doi.org/10.1016/J.INPA.2015.06.001

Ma, C., Zhang, H.H., Wang, X., 2014. Machine learning for Big Data analytics in plants. Trends Plant Sci. 19, 798–808. https://doi.org/10.1016/j.tplants.2014.08.004

Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V., 2015. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55, 263–274. https://doi.org/10.1021/ci500747n

Mahlein, A.-K., Rumpf, T., Welke, P., Dehne, H.-W., Plümer, L., Steiner, U., Oerke, E.-C., 2013. Development of spectral indices for detecting and identifying plant diseases. Remote Sens. Environ. 128, 21–30. https://doi.org/10.1016/J.RSE.2012.09.019

Mahlein, A.K., 2016. Present and Future Trends in Plant Disease Detection. Plant Dis. 100, 1–11. https://doi.org/10.1007/s13398-014-0173-7.2

Maia, F.G.M., Armesto, C., Ogoshi, C., Vieira, J.F., Maia, J.B., de Abreu, M.S., 2013. Behavior of isolated of Colletotrichum Gloeosporioides inoculated micropropagated in seedlings of coffee. Biosci. J. 29.

Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., Villa, P., Stroppiana, D., Boschetti, M., Goulart, L.R., Davis, C.E., Dandekar, A.M., 2015. Advanced methods of plant disease detection. A review. Agron. Sustain. Dev. 35, 1–25. https://doi.org/10.1007/s13593-014-0246-1

Neto, H.S., 2017. Temporal space analysis of patosystem relations with the magnesium, calcium and potassium nutrients. Universidade Federal de Lavras.

Paradela Filho, O. et al., 2001. The Colletotrichum coffee complex. Campinas Inst. Agronômico, Bol. Técnico IAC.

Pezzopane, J.R.M., Pedro Júnior, M.J., Thomaziello, R.A., Camargo, M.B.P. de, 2003.

Coffee phenological stages evaluation scale. Bragantia 62, 499–505. https://doi.org/10.1590/S0006-87052003000300015

Powers, D.M.W., Turk, C.C.R., 2012. Machine learning of natural language. Springer Science & Business Media.

Prabhakar, M., Prasad, Y.G., Thirupathi, M., Sreedevi, G., Dharajothi, B., Venkateswarlu, B., 2011. Use of ground based hyperspectral remote sensing for detection of stress in cotton caused by leafhopper (Hemiptera: Cicadellidae). Comput. Electron. Agric. 79, 189–198. https://doi.org/10.1016/j.compag.2011.09.012

Price, T. V., Gross, R., Ho Wey, J., Osborne, C.F., 1993. A Comparison of Visual and Digital Image-Processing Methods in Quantifying the Severity of Coffee Leaf Rust (Hemileia Vastatrix). Aust. J. Exp. Agric. 33, 97–101. https://doi.org/10.1071/EA9930097

Richter, R., 1996. Atmospheric correction of satellite data with haze removal including a haze/clear transition region. Comput. Geosci. 22, 675–681. https://doi.org/10.1016/0098-3004(96)00010-6

Richter, R., Schläpfer, D., 2011. Atmospheric/Topographic Correction for Satellite Imagery. In : DLR Report DLR-IB 565-02/11. DLR Rep. DLR-IB 565, 202.

Richter, R., Schläpfer, D., 2003. Atmospheric/topographic correction for satellite imagery: ATCOR-2/3 user guide, version 9.1.1, February 2017. ReSe Appl. Schläpfer 3, 270.

Rodríguez, J.D., Pérez, A., Lozano, J.A., 2013. A general framework for the statistical analysis of the sources of variance for classification error estimators. Pattern Recognit. 46, 855–864. https://doi.org/10.1016/j.patcog.2012.09.007

Rodríguez, J.D., Pérez, A., Lozano, J.A., 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rouse, J.W. 1974, 1973. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation.

Russo, D.P., Zorn, K.M., Clark, A.M., Zhu, H., Ekins, S., 2018. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. Mol. Pharm. 15, 4361–4370. https://doi.org/10.1021/acs.molpharmaceut.8b00546

Sera, G.H., Altéia, M.Z., Sera, T., Petek, M.R., Ito, D.S., 2005. Correlation among the Colletotrichum spp. incidence with some coffee agronomic traits. Bragantia 64, 435–440. https://doi.org/http://dx.doi.org/10.1590/S0006-87052005000300013.

Tan, B., Masek, J.G., Wolfe, R., Gao, F., Huang, C., Vermote, E.F., Sexton, J.O., Ederer, G., 2013. Improved forest change detection with terrain illumination corrected Landsat images. Remote Sens. Environ. 136, 469–483.

https://doi.org/10.1016/J.RSE.2013.05.013

Tucker, C.J., Grant, D.M., Dykstra, J.D., 2013. NASA's Global Orthorectified Landsat Data Set. Photogramm. Eng. Remote Sens. 70, 313–322. https://doi.org/10.14358/pers.70.3.313

Varzea, V.M.P., Rodrigues, C.J., Lewis, B.G., 2002. Distinguishing characteristics and vegetative compatibility of Colletotrichum kahawe in comparison with other related species from coffee. Plant Pathol. 51, 202–207. https://doi.org/10.1046/j.1365-3059.2002.00622.x

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens. Environ. 185, 46–56. https://doi.org/10.1016/j.rse.2016.04.008

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

West, J., Wahlen, S., McCartney, A., Ramon, H., Bravo, C., Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., Ramon, H., 2004. Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. Comput. Electron. Agric. 44, 173–188. https://doi.org/10.1016/j.compag.2004.04.003

Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. J. Inf. Sci. 44, 48–59. https://doi.org/10.1177/0165551516677946

Zhao, J., Huang, L., Huang, W., Zhang, D., Yuan, L., Zhang, J., Liang, D., 2014. Hyperspectral measurements of severity of stripe rust on individual wheat leaves. Eur. J. Plant Pathol. 139, 401–411. https://doi.org/10.1007/s10658-014-0397-6

PAPER 4 - REMOTE EVALUATION OF THE COFFEE YIELD BY MACHINE LEARNING TECHNIQUES AND SPECTRAL AGROMETEOROLOGICAL MODEL

Journal standards: *Precision Agriculture* ISSN 1573-1618

Preliminary Version

Jonathan da Rocha Miranda[a]*, Marcelo de Carvalho Alves[a], Edson Ampélio Pozza[b], Helon Santos Neto[b]

[a] Department of Agricultural Engineering at the Federal University of Lavras, University Campus, PO box: 3037, ZIP Code: 37200-000, Lavras, Minas Gerais, Brazil

[b] Department of Plant Pathology, Federal University of Lavras, University Campus, PO box: 3037, ZIP Code: 37200-000, Lavras, Minas Gerais, Brazil.

Abstract: The aim of this paper was to develop a methodology for estimating coffee yield at Landsat pixel level by machine learning techniques compared to the spectral agrometeorological. Polynomial regression in Landsat images was used to estimate radiometric values at dates when image acquisition failed, thus allowing the use of images at 16-day intervals. Landsat set of images was estimated from 09/01/2013 to 02/28/2014 and 09/01/2014 to 02/28/2015 in 16-day intervals. Random Forest modeling was performed for each set of images by using the wavelength spectral bands from blue to infrared shortwave, as well as the NDVI, NDWI and EVI spectral indices. For the spectral agrometeorological, the meteorological variables such as air temperature, evapotranspiration and NDVI and IAF indices derived from Landsat images that were used. The coefficient of determination (R^2) for the Random Forest model ($R^2 = 0.58$) was higher than for the spectral agrometeorological ($R^2 = 0.10$) to estimate coffee yields. Probably, the worst performance of the spectral agrometeorological occurred by using uniform meteorological data throughout the evaluated area, determining few spatial variations of the information used in the method. The Random Forest model, trained with yield data from two harvests was able to estimate the yield every 16 days, from September to February of the corresponding harvest year of the analysis. With a solid database of geospatial and temporal data on the yield of georeferenced in sample meshes in crops, machine learning algorithms can be used as training to determine the yield of coffee on Landsat image crops.

Keywords: Random Forest; bienniality of the coffee tree; arabica coffee; fruit yield monitoring; Landsat.

Introduction

Coffee is one of the most appreciated drinks in the world. It is estimated that more than two billions of cups are served per day, and this is constantly growing. Regarding 2019, more than 166 million 60-kg bags were consumed (ICO 2020). Coffee production has a prominent role in the world economy, involving around 500 million people who work directly or indirectly in this chain. This economic importance makes coffee one of the main commodities in the stock market, moving high daily financial volumes (Vegro and Almeida, 2019).

In the 2019 harvest, Brazil accounted for approximately 33% of world production and is considered to be the world market price leader (CONAB 2019). This economic importance of the country attracts the interest of monitoring the Brazilian harvest, with this fact being constantly analyzed by organizations, such as the National Supply Company (CONAB) and the International Coffee Organization (IOC).

The estimated coffee crop operates in both the external and domestic markets, not to mention that producers, traders and investors await the reports issued by the aforementioned organizations to plan their activities (Lima et al. 2008; Miranda et al. 2014). However, those are reports that make a predicted estimate and often present divergences between the CONAB and IOC researches. For the 2016, 2017 and 2018 harvests, the CONAB report indicated Brazilian production to be around 50, 45 and 62, while the IOC indicated a production at around 56, 52 and 64 million bags, respectively. Therefore, the search for production estimation methods is not a consolidated method and, thus, requires extensive research to ensure the reliability of these estimates.

Sampling methods to estimate yields can be based on parameters such as count of flowers, determination of the volume of green coffee or counting of ripe fruit (Schattan, 1964). However, they are methods of high transport costs, which requires complex logistics and demand too much time for these tasks to be performed.

Crop yield estimation by remote sensing is gaining visibility due to satisfactory model accuracy results (Kouadio et al. 2014; Lopresti et al. 2015; Prasad et al. 2007). The methods applied are generally based on empirical and statistical models that need to be calibrated for different agroclimatic zones on account of changes in environmental and crop conditions in different locations (Noureldin et al. 2013).

In coffee production, a spectral agrometeorological model was used by adapting the agrometeorological model of crop prediction developed by Doorenbos and Kassam (1979). The incorporation of the spectral term into the model, represented by the vegetation index, made it possible to establish relationships with agronomic variables, such as coffee tree fruits yield (Bernades et al., 2012; Rosa et al., 2011). However, most of these models use MODIS sensor

images, with moderate spatial resolution when compared to Landsat missions, but with high temporal frequency and greater territorial imaging that favor the monitoring of coffee producing regions (Brunsell et al. 2009).

A limitation in the use of Landsat for this model is due to the temporal discontinuity of the images, which may have the signal from the Earth's surface absorbed by the cloud cover at the time when images are obtained. The use of these images, however, has greater potential for detailing the conditions of the crops (Chemura et al. 2017). In this regard, the best scenario for agricultural monitoring would be to obtain images with a high level of detail in the shortest possible time, so that a model to monitor the yield of coffee tree fruits can be adopted by producers and cooperatives.

Machine learning is computational techniques capable of acquiring knowledge through the patterns by which data are presented (Singh et al. 2016). The use of machine learning has shown promising results to estimate yields with the use of satellite images for wheat (Pantazi et al. 2016), rice (Setiyono et al. 2018), barley and canola cultivation (Johnson et al. 2016). This technology model has the advantage over the spectral agrometeorological model of using information only regarding the reflectivity of the surface that is captured by the satellites. Spectral agrometeorology models use data from weather stations and MODIS sensor images, estimating the yield of coffee fruit in large areas, such as the southern mesoregion of Minas Gerais (Rosa et al. 2010) or even a single farm (Almeida et al. 2017), but were not able to assess the spatial variability of yield within a single crop.

Assuming that computer learning techniques that acquire knowledge by standards can be useful in coffee growing, the aim of this study was to evaluate machine learning techniques compared to the spectral agrometeorological model to monitor the coffee yield with greater accuracy, detail in spatial resolution and time interval.

Material and methods

Area of study and sample point selection

The area studied was located in the southern region of the state of Minas Gerais, at the municipality of Carmo do Rio Claro, in Brazil, where the coffee plantation is located, bounded by the following coordinates – latitude of 21°00'28" South of the Equator and longitude of 46°01'30" West of Greenwich (Fig 1). The coffee cultivar (Coffea arabica L.) planted was Acaiá 474/19, with a spacing of 3.6 m between rows and 0.70 m between plants, in a total area of 11 hectares. The crop was drip-irrigated with an irrigation management based on measured water demand by means of properly installed tensiometer batteries.

The sample grid was made up of georeferenced points at spacing of 40 by 40 meters, collected through a TRIMBLE 4600 LS GPS™ and Leica TC600 Total™ Station. These points were considered as a buffer of 7.2 meters of radius, corresponding to the spacing between two planting rows. The criteria to select the points was established on the condition that the buffer polygon is contained in a single pixel of the Landsat image, as recommended by Miranda et al. (2020). Resampling was performed by moving the point to the closest position to the center of the pixel in which it is contained in order to reorganize a mesh structure coinciding with those of Landsat images.

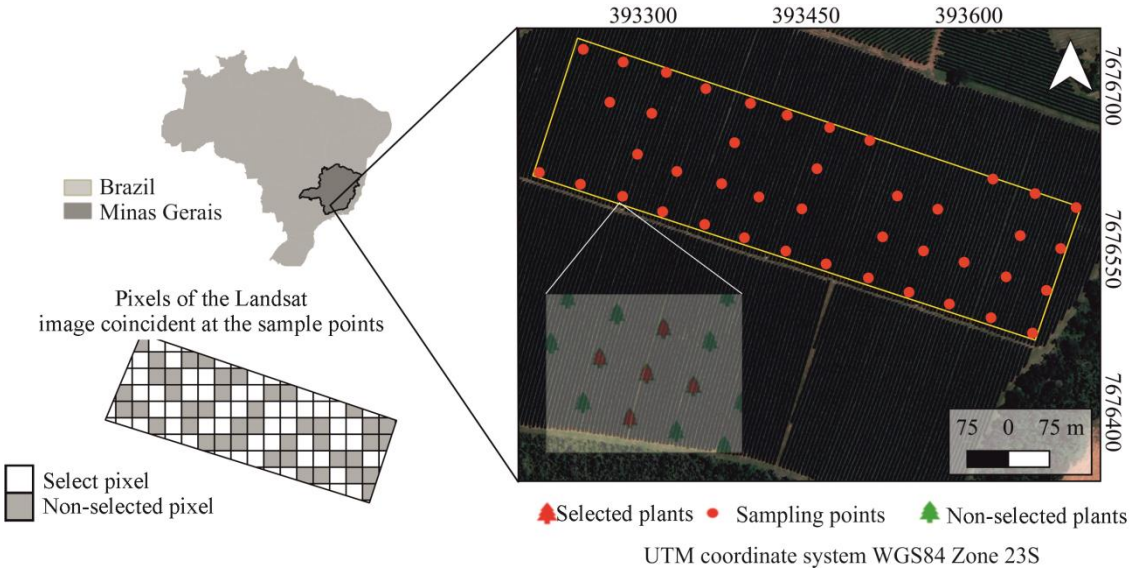

Fig 1. Spatial location of the area under study, highlighting the distribution of the sample set, emphasizing the 5 coffee plants selected in each sample.

The yield for the years of 2013, 2014 and 2015 was achieved by means of manual harvesting in each of the 43 sample points. Every volume collected per plant was inserted in a graduated container. The yield of coffee was measured in liters per plant and later converted into 60-kg bags per hectare (bags/ha).

Orbital data acquisition

Google Earth Engine Code was used to acquire spectral data from the Landsat 7 ETM+ satellite and 8 OLI reflectance surface product, which indicates that the images were corrected for atmospheric effects using the 6SV (Table 1). Google Earth Engine is a virtual programming platform in Java or Python in which it has access to Google's geographic database, which contains the collection of images from several surface monitoring satellites, such as Landsat, Sentinel, Terra, Aqua among others (Gorelick et al. 2017).

The programming routine in JavaScript consisted in extracting the radiometric values from the images to the sample points under the condition that all radiometric values are cloud-free and with low aerosol interference. This selection criterion was performed by observing the values in the pixel quality band, which is a product originating from the atmospheric correction that classifies the pixel according to the atmospheric condition present in the location, being the value of 66 for Landsat 7 ETM+ and 322 for Landsat 8 OLI the best quality pixel class.

Table 1. Wavelength for the Landsat 7 ETM+ and 8 OLI satellite spectral bands, which were used as input for the model to estimate coffee fruit yields

Bands	Spectral resolution Wavelength (mm)	
	Landsat 8 OLI	Landsat 7 ETM+
Blue	0.45 - 0.51	0.45 - 0.52
Green	0.53 - 0.59	0.50 - 0.60
Red	0.64 - 0.67	0.63 - 0.69
NIR	0.85 - 0.88	0.76 - 0.90
SWIR 1	1.57 -1.65	1.55 - 1.75
SWIR 2	2.11 - 2.29	2.08 - 2.35

The radiometric values for the period of September 2013 to February 2014 and September 2014 to February 2015 were extracted, which correspond to the phenological phase of the beginning of flowering and fruit ripening (Pezzopane et al. 2003) (Fig 2).

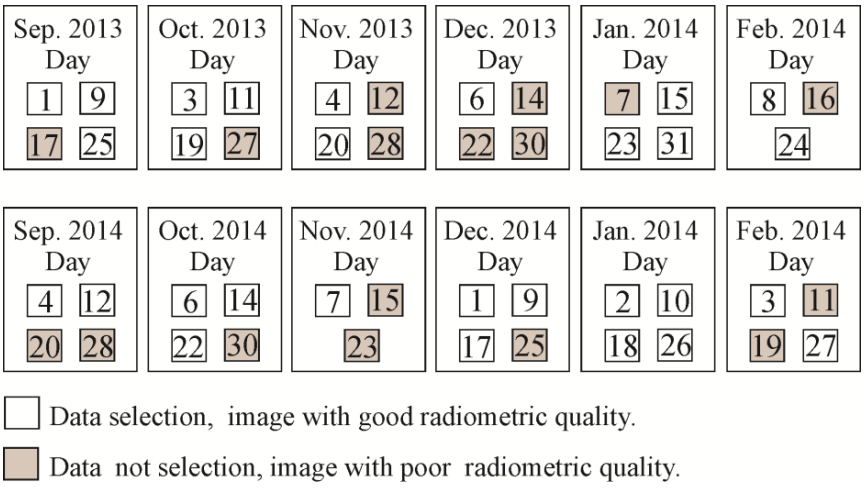

Fig 2. Dates of selected images highlighting those when there was good quality in radiometric value in Carmo do Rio Claro, MG.

The radiance values were converted into surface reflectance by correcting atmospheric effects by using the 6SV method (Vermote et al. 2016). The vegetation index was calculated by Normalized Difference Vegetation Index (NDVI) (Rouse, 1973), Enhanced Vegetation Index (EVI) (Justice et al., 1998) and Normalized Difference Water Index (NDWI) (Gao 1996). These indices were selected because they address at least one divergent band among them (Equations 1, 2 and 3).

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \quad (1),$$

$$EVI = 2.5 * \frac{\rho_{nir} - \rho_{red}}{1 + \rho_{nir} + 6 * \rho_{red} - 7.5 * \rho_{blue}} \quad (2),$$

$$NDWI = \frac{\rho_{nir} - \rho_{swir1}}{\rho_{nir} + \rho_{swir1}} \quad (3),$$

where ρ_{nir} is the Near Infrared surface reflectance band; ρ_{red} is the Red surface reflectance band; ρ_{swir1} is the Shortwave infrared surface reflectance band; ρ_{blue} is the Blue surface reflectance band.

Polynomial Regression Model

The polynomial regression (Equation 4) was used to adjust the curve behavior of radiometric values over time for the spectrum bands from blue to shortwave infrared and NDVI, EVI and NDWI spectral indices. The dates were converted into continuous days after September 1st. Thus, each regressive model had, as a dependent variable, the reflectance value of the corresponding band and, regarding the independent variable, the dates were expressed in continuous days.

$$y = \alpha_0 + \alpha_1.x + \alpha_2 x^2 + \alpha_3.x^3 + \alpha_4.x^4 \quad (4),$$

where y is the estimated spectral value; α, the coefficients adjusted by the method of least squares; and x, the contiunous day after September 1st.

With the adjusted regression model, the daily interval radiometric values were estimated for the use of the spectral agrometeorological model and, at a 16-day interval, to be used in the Random Forest regression.

Spectral agrometeorological model

The yield estimation model proposed by Doorenbos and Kassam (1979), adapted to the additive model of penalizing yield of the previous year due to the biennial nature of the coffee crop, according to Santos and Camargo (2006) (Equation 5) is as follows:

$$Ye = \sum_{i=1}^{353}\left[Yp_i\left(1-\left[1-kyo\left(\frac{Yaa}{Ypi}\right)\right].\left[1-kyi\left(1-\frac{ETr}{ETc}\right)\right]\right)\right] \quad (5),$$

where Ye is the estimated final yield (kg·ha^{-1}); Ypi, the potential partial yield of the crop (kg·ha^{-1}); kyi is the water penalty coefficient for each phenological adimensional period; ETr is the actual evapotranspiration (mm·d^{-1}); ETc, the evapotranspiration of culture (mm. d^{-1}); Yaa, the previous year's yield (kg. ha^{-1}); kyo, the penalty coefficient due to yield of the previous year; and i, the days of the images.

Potential crop evapotranspiration (ETp) was determined by using the Penman-Monteith equation (Allen et al., 1998). Air temperature data [minimum, average and maximum (°C)], wind speed at 2 meters high (m.s^{-1}) and relative air humidity (%) were used.

For the water penalty coefficient (ky), the reference values defined by Santos and Camargo (2006), which carried out the calibration at farm level in crops in the state of São Paulo for each phenological phase of the coffee tree.

In the penalty coefficient, determined by the yield of the previous year (kyo), the value of 0.5 proposed by Almeida et al. (2017) was used for the same values of ky, as proposed by Santos and Camargo (2006), for the calibration of the kyo coefficient.

The potential yield was determined by incorporating the spectral term into the agrometeorological model for the estimates of biophysical parameters of vegetation proposed by Rizzi and Rudorff (2007). In this case, we used the model suggested by Doorenbos and Kassam (1979), although with the addition of the Leaf Area Index (LAI) for the calculation of potential yield (Equation 6).

$$Yp = cL.cN.cH.G.Yo \quad (6),$$

where Yp is the yield potential (kg.d.ha^{-1}); cL, the growth compensation factor and leaf area; cN, the dry matter production factor; cH, the harvesting factor; G, the number of days between images; and Yo, the gross dry matter production (kg.day.ha^{-1}).

To calculate the growth compensation factor (cL) we used the model of Sugawara (2002), determined on the basis of data from Doorenbos and Kassam (1979) (Equation 7).

$$cL = 0{,}515 - e^{[-0{,}664-(0{,}515.LAI)]} \quad (7),$$

where LAI is the Leaf Area Index.

Rizzi and Rudorff (2007) recommend that the spectral variable (LAI) required by the model to be obtained from the NDVI vegetation index. In this sense, the NDVI images were first transformed into ground cover fraction images (Choudhury et al., 1994) (Equation 8).

$$F_{cor} = 1 - \left(\frac{NDVI_{MAX}-NDVI}{NDVI_{MAX}-NDVI_{MIN}}\right)^{0,6} \quad (8),$$

where F_{cor} is the fraction of soil covered by the crop; $NDVI_{max}$ and $NDVI_{min}$ are the maximum and minimum values of the vegetation index of the image; NDVI is the value of NDVI of each pixel of the image.

The LAI was, thus, obtained through the following relationship (Norman et al. 2003) (Equation 9).

$$IAF = -2\ln(1 - F_{cor})$$

(9).

The gross dry matter production, Yo (kg.day.ha^{-1}), was obtained by applying the concept of Wit (1965), which relates the fraction of the day when the sky is cloudy to the gross dry matter production rate for a standard crop depending on the crop and air temperature.

The yield factor (cH) is related to the harvested part in relation to the total dry mass of the coffee trees. For this coefficient, the values used were defined by Almeida et al. (2017). The values adopted were 0.19 and 0.10 for the high and low yield years, respectively.

The liquid dry matter production factor (cN) is related to the energy requirement for internal development processes (respiration), which depends on the temperature of the region. For the average temperature condition below 20°C, the factor adopted was equal to 0.6 and, for the average temperature condition above 20°C, the factor was equal to 0.5 (Doorenbos and Kassam 1979).

The number of days between images (G) was done on a daily basis because the values of LAI were estimated using the polynomial equation.

Yield estimation by the Random Forest regressive model

The yield estimate was calculated in a 16-day follow-up, from September 1st onwards. Polynomial regression was used to estimate radiometric values on the selected dates. The regressive model was used due to image discontinuity on the desired dates and to standardize the evaluation dates for the years of 2014 and 2015.

The input variables were the Landsat spectral bands and the NDVI, EVI and NDWI vegetation indices. Each training set was conducted for each evaluation date individually and, thus, the best date that allows an estimate of yield was established.

The Random Forest (RF) regression algorithm was used (Breiman 2001), adopting, as a dependent variable, the yield, and, in respect to the independent variable, the areas below the curve of each one. The Random Forest regression technique consists of a combination of decision trees that act independently from each other (Mutanga et al. 2012). A total of 500 decision trees were used because the errors stabilize before this number of classification trees is reached according to Lawrence et al. (2006). The ramification criterion was defined by the

lowest Mean Absolute Error (MAE), a value established in the training to decrease the randomness of the classifier. The contribution analysis of the variables in the decrease of the Mean Absolute Error (MAE) was extracted, being established by the degree of importance given in percentage.

The Random Forest model was validated using the leave-one-out method (Cawley and Talbot 2003). This method consists of repeating the validation process point by point, being that in each process a single point is taken from the training set, and iteration is carried out until all the points are predicted.

Evaluation of yield forecasting models

The yield estimation by Random Forest and the agrometeorological spectral model (Fig 3) were compared by metrics of evaluation, coefficient of determination (R^2), pearson correlation (r), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

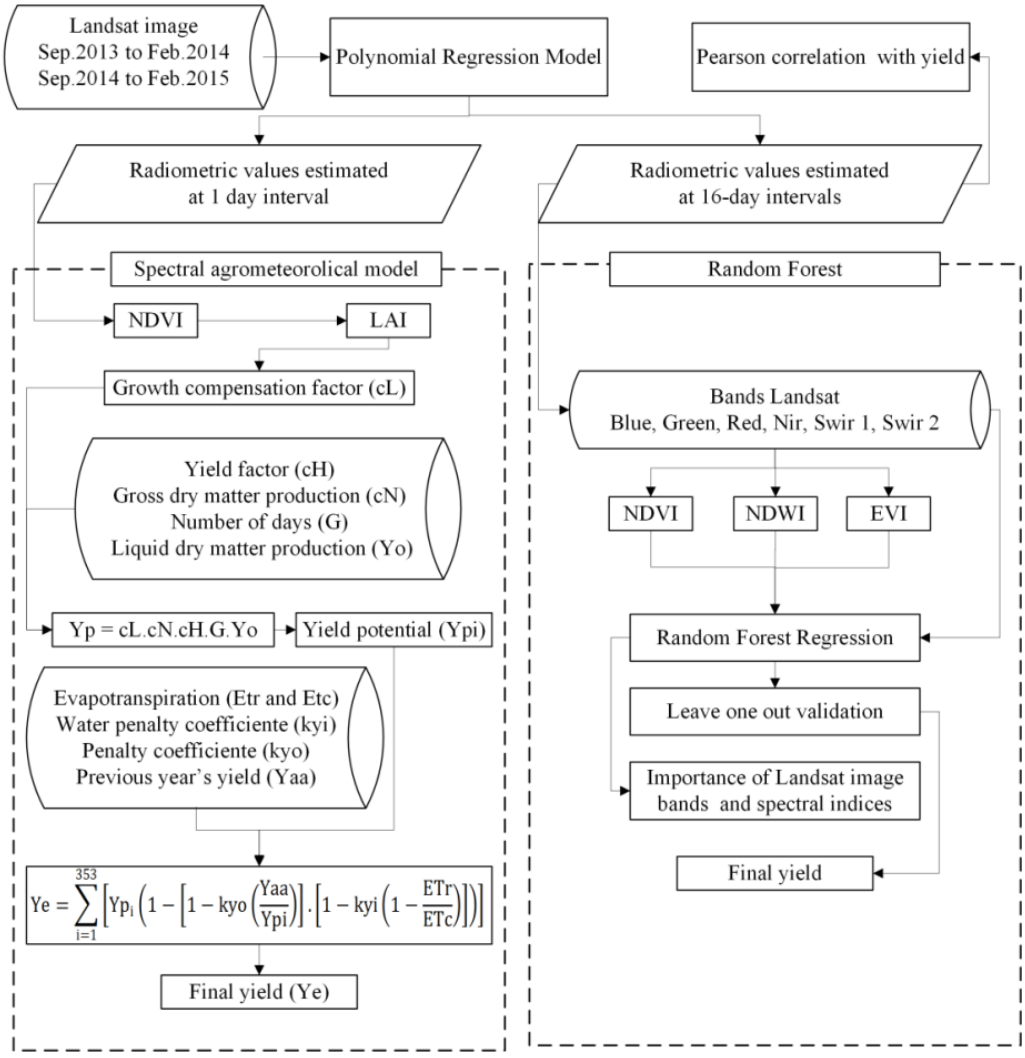

Fig 3. Flowchart of procedures adopted for estimating coffee yield by the Random Forest and spectral agrometeorological model.

Results and discussion

The yield in 2014 was higher than in 2013 and 2015, which can be considered as a reflection of the biennial nature of the crop (Fig 4 and 5). The difference in coffee yield is due to the peculiarity of its phisiology and vegetative grown. For years of high yield, most photo similarities are carried to the fruit, resulting in low vegetative development, which will culminate in lower yields in the next crop (Bernardes et al. 2012).

The effects on low yields, such as plant health of the coffee tree, may be accentuated by other factors (Pozza et al. 2010; Varzea et al. 2002). Avelino et al. (2006) reported that with high coffee production, rust intensity is more severe, resulting in high leaf fall after harvesting and, consequently, lower yields in the following crop.

The presence of the Colletotrichum sp. complex was detected in this crop, causing necrosis in the fruits in the 2014 crop (Miranda et al. 2020), but even though it achieved the higher yield compared to 2013 (Santos Neto, 2017). Silva et al. (2019) described, in the crop of the years of 2013 and 2014, the incidence of rust, but it was relatively low, between 5 and 10%, respectively.

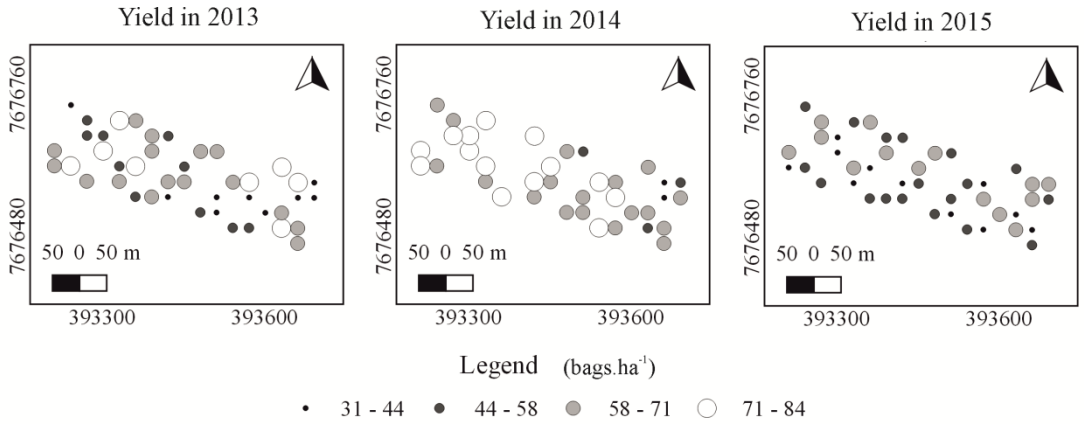

Fig 4. Spatial distribution of yield in quintiles for the 2014 and 2015 harvesting.

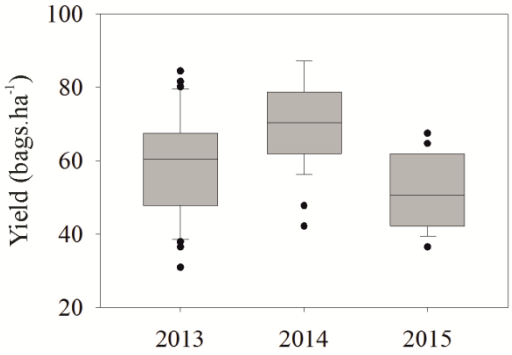

Fig 5. Boxplot of yields for the 2013, 2014 and 2015 harvesting.

As for the climatic conditions in the period assessed (Fig 6), the rainfall corresponding to the 2014 crop (period of September 2013 to June 2014) was lower than the hydric need of the crop [1,600 mm according to Camargo (2010)]. However, due to drip irrigation management, it tends not to cause water defict in the production. The temperature was within the ideal range for optimal crop performance. According to Laderach et al. (2011), the temperature that sets the range varies from 18 to 23°C.

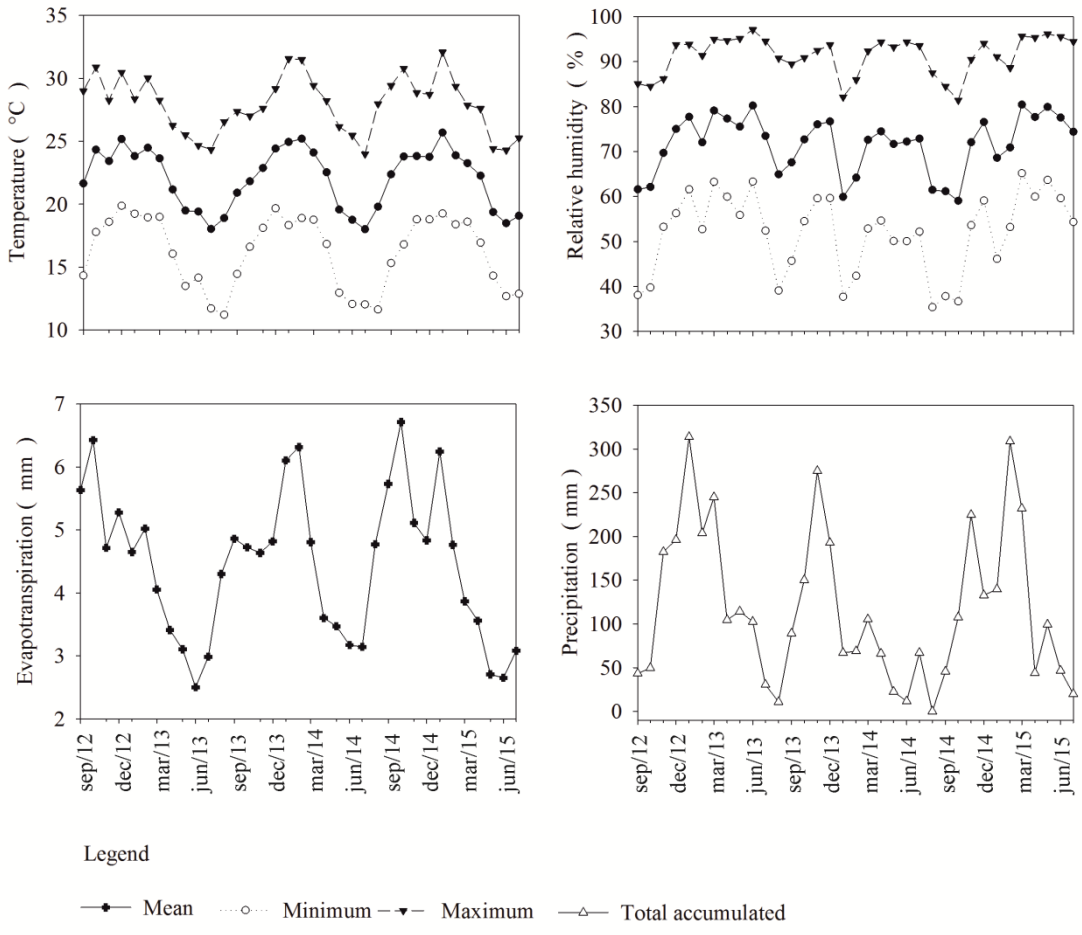

Fig 6. Climate conditions from September 2012 to June 2015 for the variables temperature, relative humidity, evapotranspiration and precipitation.

The coefficient of determination (R^2) corresponding to the period of 2013/2014 was lower in relation to the period of 2014/2015 for the polynomial function regressive model (Fig 7). The low rainfall in the 2013/2014 rainy season may have specifically influenced the Swir 1 band, which did not present the behavior of the time series characterized by the polynomial for the 2013/2014 crop year. The noise caused by the atmospheric effects that may interfere in the prediction of radiometric values should also be considered, since it does not characterize the reflectivity of the surface. According to Tan et al. (2013), even with atmospheric correction, images are susceptible to noise, which interferes in reflectance, such as the topographic effects that, in a bumpy area, cause image overlap, and this effect tends not to be corrected by the 6SV atmospheric correction method.

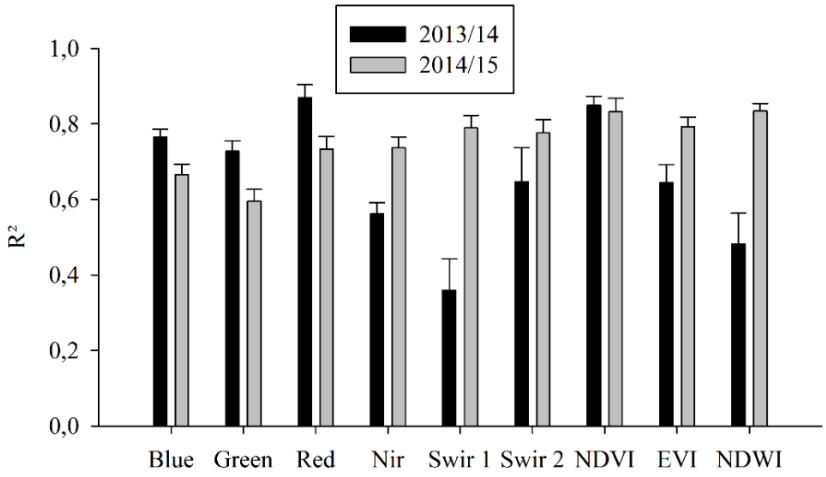

Fig 7. Mean values of coefficient of determination (R^2) with standard deviation of the polynomial regression model for the Landsat 7 ETM+ and 8 OLI image bands and spectral indices.

The correlations between radiometric values and coffee yield were significant at 5% for the period of 2013/2014, mainly in the Swir 1 band, which had significant correlations over the whole period analyzed. The same analyses performed with the radiometric values estimated by polynomial regression were also significant and stable with regard to the differences in the value of the correlation between the dates (Fig 8). However, in the following period of 2014/2015 there was not the same behavior, and few dates obtained significant correlations, either using the images or the estimate by the polynomial regression model. The 2014/2015 crop year was a period of low yield, and, due to the effects of bienniality, there was, at this time, more leaf

formation and less fruit formation. So, no pattern was determined that relates radiometric values and yield.

Fig 8. Pearson correlation between Landsat 7 ETM+ and 8 OLI bands and spectral indices with yield using the current and estimated polynomial images for the crop of years 2013/14 and 2014/15.

By utilizing the polynomial model to estimate the radiometric values of the Landsat images, it was possible to set the same dates for the 2013/2014 and 2014/2015 crop years and, thus, perform a joint analysis of the correlation of the images with the yield in two years. Image correlations with yield were significant at 5% practically throughout the period and for all bands (Fig 9). Yield in relation to spectral behavior may depend on more than one crop/year in order to explore the effects of bienniality, as in the case of spectral agrometeorological models, which use the previous year's yield for the later estimate (Almeida et al. 2017; Bernardes et al. 2012; Picini 1998; Rosa et al. 2010; Silva et al. 2011). According to Camargo and Camargo (2005),

understanding the spectral behavior of the coffee tree depends on more than one year of assessment, as the crop takes two years to complete its phenological cycle.

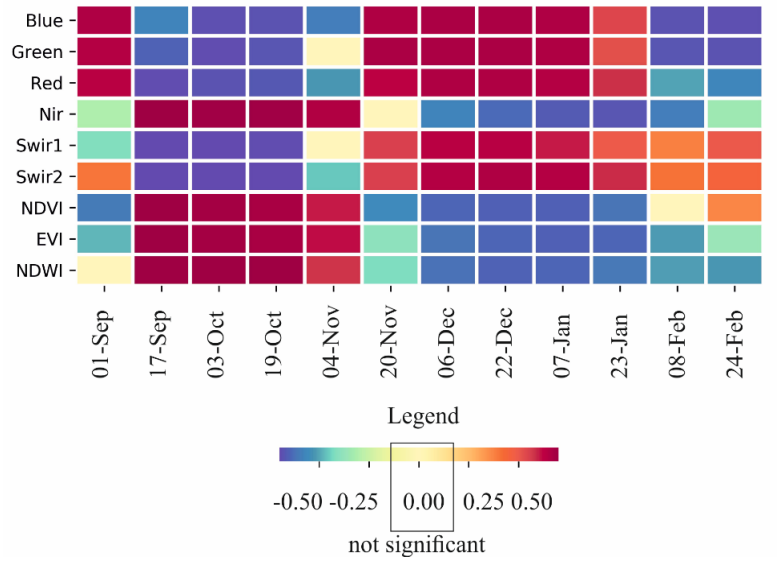

Fig 9. Pearson correlation of Landsat images estimated by polynomial model and coffee fruit yield for joint analysis of two crop years – 2013/2014 and 2014/2015.

The Random Forest regression remained above 0.40 R^2 for the whole period under analysis, and at 60 calendar days, which corresponds to the month of November (Fig 10), being obtained the best estimate (R^2 0.59 and MAE 7.6 bags.ha^{-1}). Even with relatively low R^2, obtaining a yield estimate continuously over time can be used as a crop management tool in the targeted application of agricultural inputs according to production potential. According to Martinelli et al. (2015), crop phytosanitary monitoring by means of remote sensing techniques has an essential function in the management of rapid decision-making.

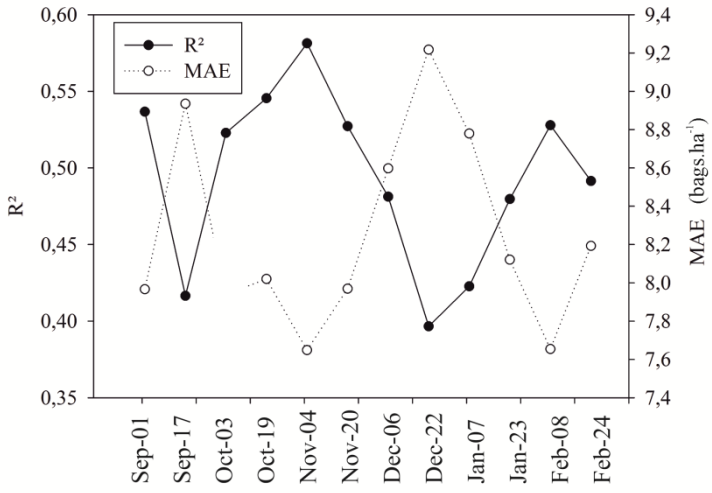

Fig 10. Coefficient of determination (R^2) and Mean Absolute Error (MAE) of Random Forest fruit yield estimation in Landsat images in a two-year joint analysis estimated by the polynomial model.

In February, the blue band was the most important, being a period that corresponds to the phenological phase of fruit granulation (Fig 11). The spectral blue band has a relation with the decline in the amount and composition of chlorophyll foliar (Feng et al. 2016; Ustin et al. 2009). The blue band was also used for nutritional stress (Mutanga and Skidmore 2004) and other plant health disorders (Carter and Knapp 2001).

The Swir 1 band was more important from September to October, which corresponded to the driest period of the year (Fig 11). The irrigation of the crop was managed according to the water tension in the soil by means of tensiometers – if the soil type is divergent along the crop, the tensiometers installed and concentrated in a single region may not represent the entire extension of the area. In this case, there may be water deficit somewhere in the crop, and, in the dry season, this factor may be more evident in order to be characterized by the Swir 1 band and, thus, obtain greater weight in the yield estimate.

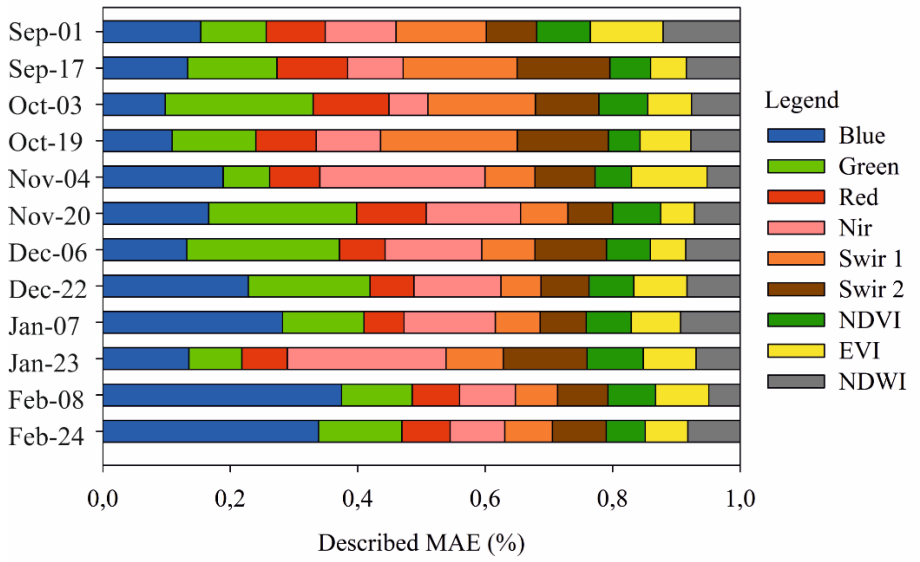

Figure 11. Importance of Landsat image bands and spectral indices for estimating coffee fruit yields over time.

As for the distribution of predicted and expected values for the Random Forest, it was considered the mean forecast for the entire period analyzed, while regarding the spectral agrometeorological, it was considered the final sum of the production potential (Fig 12). Random Forest was able to characterize that the 2014 crop year was higher than 2015, with a completely separable group in the distribution of the predicted values.

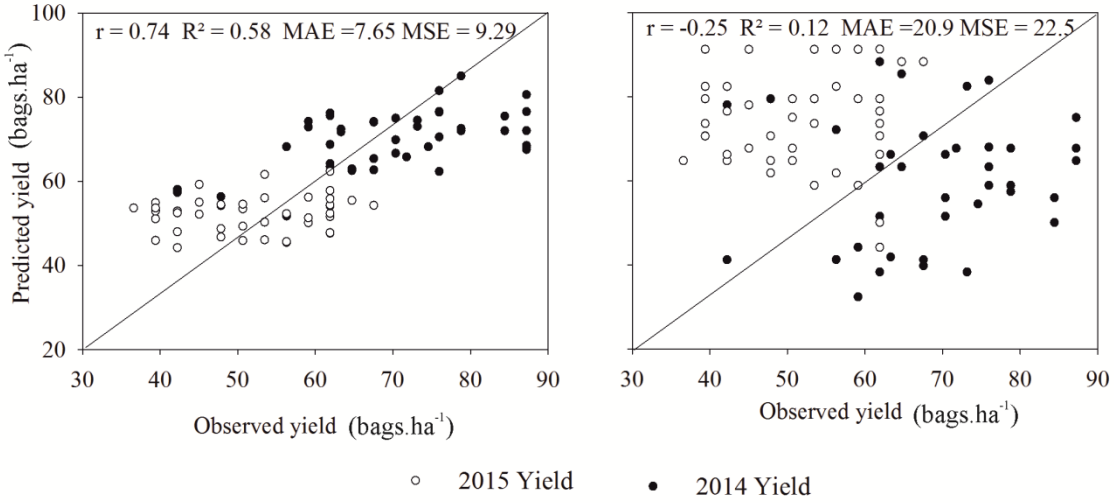

Fig 12. Prediction of coffee tree fruit yield by Random Forest and spectral agrometeorological model.

However, the spectral agrometeorological was the opposite, given that the predicted values for the year of 2015 were higher than the year of 2014. The spectral agrometeorological model requires climate variables to impose the water deficit and penalize potential yield, however, these variables had no spatial variation. It should be noted that the previous year's yield consists in a weighting factor for the estimation of the yield by the spectral agrometeorological because there is no spatial variation of the hydric penalty. Consequently, this yield had a greater impact on the predictions in order to underestimate the 2014 yield and overestimate that one referred to 2015.

The yield forecast models for small scales of property level obtained consistent results with the spatial distribution similar to in situ data. For the spectral-agrometeorological model for coffee yield prediction, which used the survey of the Brazilian Institute of Geography and Statistics (IBGE), Rosa et al. (2010), when estimating the crop over 5 yields, obtained the R^2 of 0.72 for the set of crops in southern Minas Gerais. Almeida et al. (2017), when evaluating crops in northwest Minas Gerais, obtained the R^2 ranging from 0.79 to 0.95 by using the spectral-agrometeorological.

In the agrometeorological model, it also presented estimates with a small amount of errors in relation to IBGE values. According to Victorino et al. (2016), the agrometeorological model for the coffee plantations located in Lavras, in southern Minas Gerais, obtained a maximum correlation of 0.89. By using the in situ crop, Aparecido and Rolim (2018), when estimating yield by multiple regression model with meteorological variables for the crops of the Regional Cooperative of Coffee Growers in Guaxupé – Cooxupé, in a time series of 18 years, obtained absolute percentage errors $\leq 2.9\%$.

The Random Forest model, at the 900 m² Landsat pixel level, was used to obtain results with a spatial distribution similar to that one found in situ and with percentage errors less than 13% for most samples (Fig 13 and 14). The spectral agrometeorological model presented a behavior of spatial distribution similar to the previous year's yield. Silva et al. (2011) used the agrometeorological model to predict the coffee crop in a sample mesh of 100 points, and there was an overestimation of the crop in an average of 27%. The authors observed that possible agricultural operations and uncontrollable variables in the evaluations, such as nutrition and phytotechnical factors, may have been affected in isolated points, which corroborated the errors in harvest estimation.

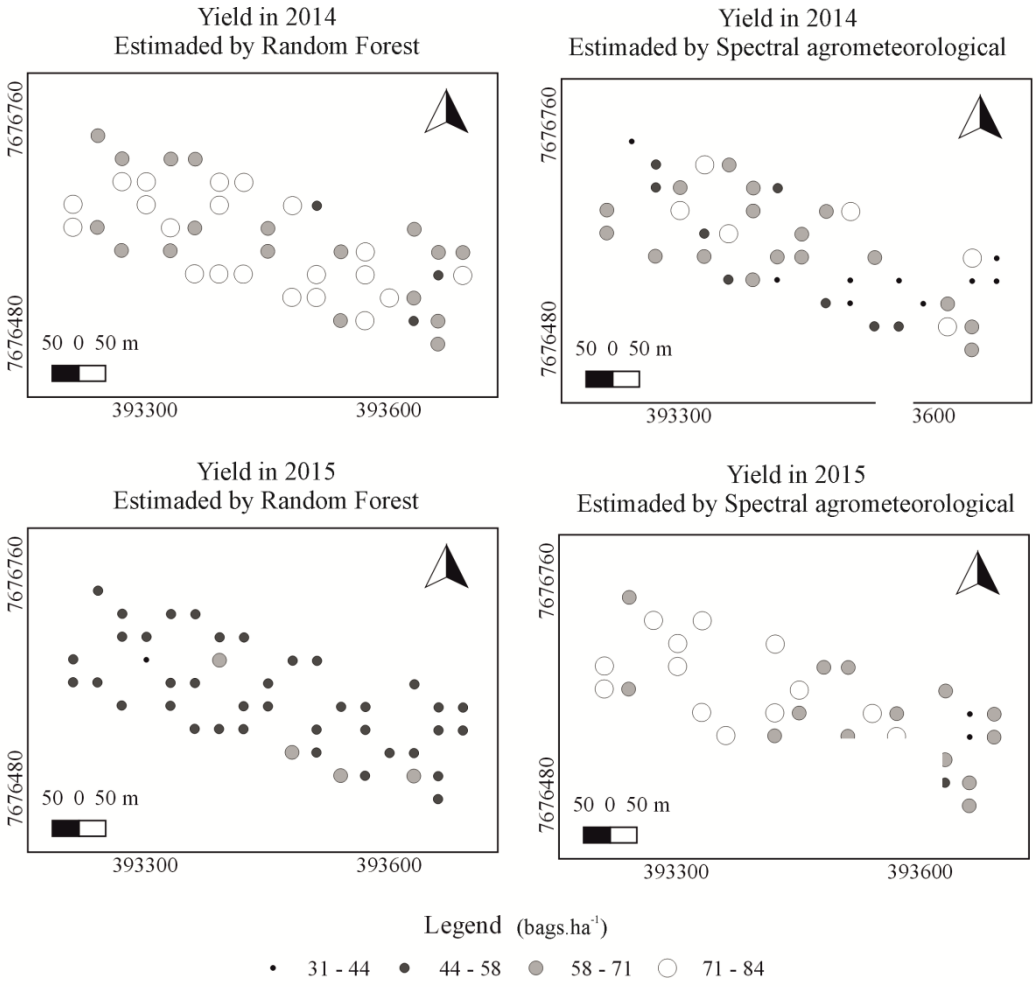

Fig 13. Spatial distribution of coffee yield predicted by Random Forest and the spectral agrometeorological model.

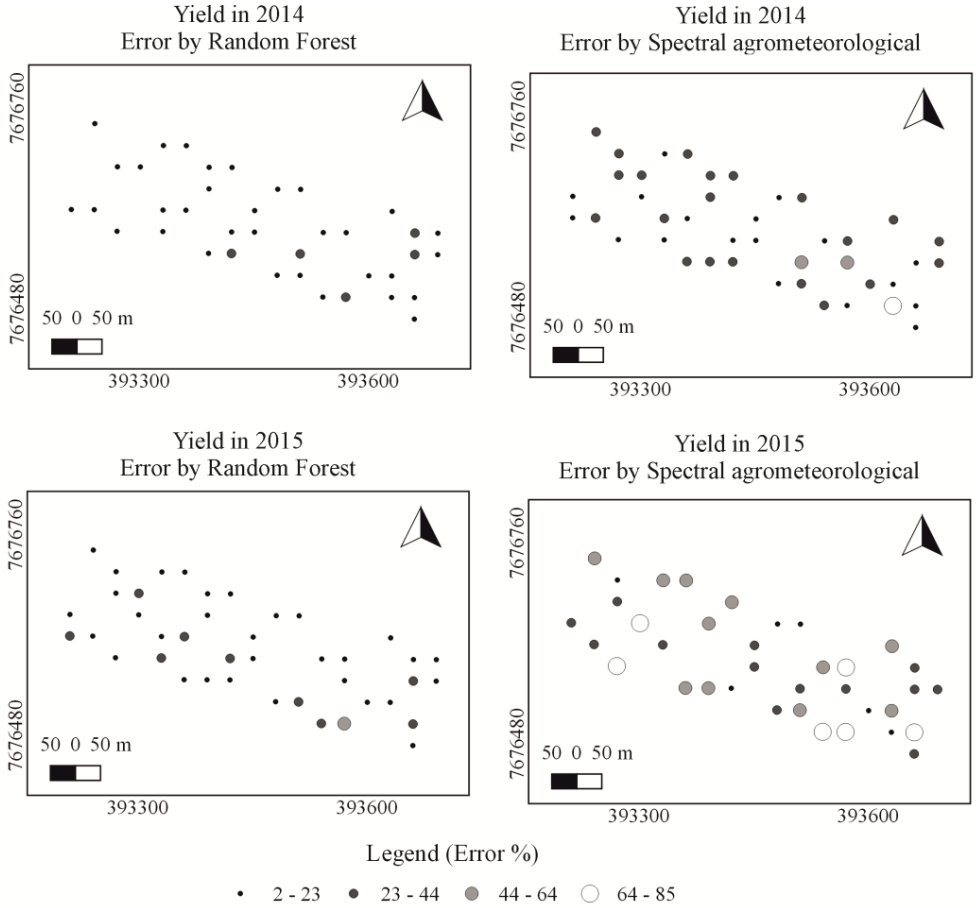

Fig 14. Error map of coffee yield estimated by Random Forest and the spectral agrometeorological model.

Conclusion

Random Forest coffee yield prediction using Landsat 7 ETM+ and Landsat 8 OLI images was able to estimate yield with significant hits for a pixel-level estimate. The machine learning can be improved by being subjected to more years of analysis for the Random Forest training base. The lacks of pixel-level meteorological data from Landsat, as well as the one regarded to the spectral agrometeorological were unable to estimate pixel-level yield. By having a historical database of coffee fruit yields for training, it was possible to obtain good performance for predicting and monitoring coffee yield in coffee plantations.

Acknowledgements The authors wish to thank (i) the Department of Agricultural Engineering of the Federal University of Lavras (UFLA) for providing office space and infrastructure to achieve the results obtained in this article, as well as (ii) the Minas Gerais Research Funding Foundation (FAPEMIG), (iii) Coordination of the Improvement of Higher Education Personnel

(CAPES), (iv) National Council for Scientific and Technological Development (CNPq) and (v) Padre Victor Farm and Boa Esperança Farm, in the city of Carmo do Rio Claro, state of Minas Gerais for the support in field experiments.

References

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. (1998). *Crop evapotranspiration – Guidelines for computing crop water requirements. – FAO Irrigation and drainage paper 56 / Food and Agriculture Organization of the United Nations. Irrigation and Drainage FAO, Rome.* (Vol. 156).

Almeida, T. S., Sediyama, G. C., & De Alencar, L. P. (2017). Estimation of the Yield of Irrigated Coffee Growers by the Spectral Agroecological Zone Method . *Revista Engenharia Na Agricultura - Reveng*, *25*(1), 1–11. https://doi.org/10.13083/reveng.v25i1.727

Aparecido, L. E. de O., & Rolim, G. de S. (2018). Forecasting of the annual yield of Arabic coffee using water deficiency. *Pesquisa Agropecuaria Brasileira*, *53*(12), 1299–1310. https://doi.org/10.1590/S0100-204X2018001200002

Avelino, J., Zelaya, H., Merlo, A., Pineda, A., Ordóñez, M., & Savary, S. (2006). The intensity of a coffee rust epidemic is dependent on production situations. *Ecological modelling*, *197*(3–4), 431–447. https://doi.org/10.1016/j.ecolmodel.2006.03.013

Bernardes, T., Alves Moreira, M., Adami, M., & Friedrich Theodor Rudorff, B. (2012). Monitoring biennial bearing effect on coffee yield using MODIS remote sensing imagery. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *4*(9), 3760–3763. https://doi.org/10.1109/IGARSS.2012.6350499

Breiman, L. (2001). RANDOM FORESTS Leo. *Machine learning*, *45*(1), 1–33. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brunsell, N. A., Pontes, P. P. B., & Lamparelli, R. A. C. (2009). Remotely sensed phenology of coffee and its relationship to yield. *GIScience & Remote Sensing*, *46*(3), 289–304. https://doi.org/10.2747/1548-1603.46.3.289

Camargo, â. P. De, & Camargo, m. B. P. De. (2001). Definition and layout of phenological phases of Arabica coffee in tropical conditions in Brazil. *Bragantia*, *60*(1), 65–68. https://doi.org/10.1590/s0006-87052001000100008

Camargo, M. B. P. de. (2010). The impact of climatic variability and climate change on arabic coffee crop in Brazil. *Bragantia*, *69*(1), 239–247. https://doi.org/10.1590/S0006-87052010000100030

Carter, G. A., & Knapp, A. K. (2001). Leaf optical properties in higher plants: linking spectral

characteristics to stress and chlorophyll concentration. *American journal of botany*, *88*(4), 677–684. https://doi.org/10.2307/2657068

Cawley, G. C., & Talbot, N. L. C. (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, *36*(11), 2585–2592. https://doi.org/10.1016/S0031-3203(03)00136-5

Chemura, A., Mutanga, O., & Dube, T. (2017). Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. *Precision Agriculture*, *18*(5), 859–881. https://doi.org/10.1007/s11119-016-9495-0

Choudhury, B. J., Ahmed, N. U., Idso, S. B., Reginato, R. J., & Daughtry, C. S. T. (1994). Relations between evaporation coefficients and vegetation indices studied by model simulations. *Remote Sensing of Environment*, *50*(1), 1–17. https://doi.org/10.1016/0034-4257(94)90090-6

CONAB. (2019, September). Acompanhamento da Safra Brasileira. *Companhia Nacional de Abastecimento*. https://www.conab.gov.br/info-agro/safras/cafe/boletim-da-safra-de-cafe.

Doorenbos, J., & Kassam, A. H. (1979). Yield response to water. *Irrigation and drainage paper*, *33*, 257.

Feng, W., Shen, W., He, L., & Duan, J. (2016). Improved remote sensing detection of wheat powdery mildew using dual-green vegetation indices, 608–627. https://doi.org/10.1007/s11119-016-9440-2

Gao, B.-C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, *58*(3), 257–266. https://doi.org/https://doi.org/10.1016/S0034-4257(96)00067-3

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Remote Sensing of Environment Google Earth Engine : Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*(2016), 18–27. https://doi.org/10.1016/j.rse.2017.06.031

ICO, I. C. O. (2020). Arabica Group Prices Fall in May While Volatility Subsides. http://www.ico.org/pt/new_historical_p.asp?section=Estat%EDstica

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and forest meteorology*, *218*, 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003

Justice, C. O., Vermote, E., Townshend, J. R. G., Defries, R., Roy, D. P., Hall, D. K., et al. (1998). The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, *36*(4), 1228–1249. https://doi.org/10.1109/36.701075

Kouadio, L., Newlands, N., Davidson, A., Zhang, Y., & Chipanshi, A. (2014). Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale. *Remote Sensing*, *6*(10), 10193–10214. https://doi.org/10.3390/rs61010193

Laderach, P., Lundy, M., Jarvis, A., Ramirez, J., Portilla, E. P., Schepp, K., & Eitzinger, A. (2011). Predicted impact of climate change on coffee supply chains. In *The economic, social and political elements of climate change* (pp. 703–723). Springer. https://doi.org/10.1007/978-3-642-14776-0_42

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*(3), 356–362. https://doi.org/10.1016/j.rse.2005.10.014

Lima, L. A., Custódio, A. A. de P., & Gomes, N. M. (2008). Coffee yield and production during the initial five harvests under irrigation with center pivot in Lavras, MG. *Ciencia e Agrotecnologia*, *32*(6), 1832–1842. https://doi.org/10.1590/S1413-70542008000600023

Lopresti, M. F., Di Bella, C. M., & Degioanni, A. J. (2015). Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. *Information Processing in Agriculture*, *2*(2), 73–84. https://doi.org/10.1016/J.INPA.2015.06.001

Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., et al. (2015). Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development*, *35*(1), 1–25. https://doi.org/10.1007/s13593-014-0246-1

Miranda, J. da R., de Carvalho Alves, M., Pozza, E. A., & Neto, H. S. (2020). Detection of coffee berry necrosis by digital image processing of landsat 8 oli satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, *85*, 101983. https://doi.org/10.1016/j.jag.2019.101983

Miranda, J. M., Reinato, R. A. O., & Silva, A. B. da. (2014). Mathematical model for predicting coffee yield. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *18*(4), 353–361. https://doi.org/10.1590/S1415-43662014000400001

Mutanga, O., Adam, E., & Azong, M. (2012). International Journal of Applied Earth Observation and Geoinformation High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observations and Geoinformation*, *18*, 399–406. https://doi.org/10.1016/j.jag.2012.03.012

Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*, *25*(19), 3999–4014. https://doi.org/10.1080/01431160310001654923

Neto, H. S. (2017). Space-time analysis of the patosystem relationship with the mineral nutrients magnesium , calcium and potassium. *Thesis*. Universidade Federal de Lavras. Retrieved from http://repositorio.ufla.br/jspui/handle/1/34237

Norman, J. M., Anderson, M. C., Kustas, W. P., French, A. N., Mecikalski, J., Torn, R., et al. (2003). Remote sensing of evapotranspiration for precision-farming applications. *International Geoscience and Remote Sensing Symposium*, 21–25.

Noureldin, N. A., Aboelghar, M. A., Saudy, H. S., & Ali, A. M. (2013). Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science*, *16*(1), 125–131. https://doi.org/10.1016/j.ejrs.2013.04.005

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65. https://doi.org/10.1016/j.compag.2015.11.018

Pezzopane, J. R. M., Pedro Júnior, M. J., Thomaziello, R. A., & Camargo, M. B. P. de. (2003). Scale for evaluation of phenological stages of Arabica coffee. *Bragantia*, *62*(3), 499–505. https://doi.org/10.1590/S0006-87052003000300015

Picini, A. G. (1998). Development and testing of agrometeorological models for estimating coffee tree (Coffea arabica L.) yields from monitoring soil water availability. 1998. *PhD Thesis.* Luiz de Queiroz College of Agriculture, University of São Paulo.

Pozza, E. A., Carvalho, V. L. de, & Chalfoun, S. M. (2010). Symptoms of Injury Caused by Diseases in Coffee Growers. *Semiologia do Cafeeiro: Sintomas de Desordens Nutricionais, Fitossanitárias e Fisiológicas*, 67–106.

Prasad, A. K., Singh, R. P., Tare, V., & Kafatos, M. (2007). International Journal of Remote Sensing Use of vegetation index and meteorological parameters for the prediction of crop yield in India Use of vegetation index and meteorological parameters for the prediction of crop yield in India. *International Journal of Remote Sensing*, *28*(23), 5207–5235. https://doi.org/10.1080/01431160601105843

Rizzi, R., & Rudorff, B. F. T. (2007). MODIS sensor images associated with an agronomic model to estimate soybean yields. *Pesquisa Agropecuária Brasileira*, *42*(1), 73–80. http://dx.doi.org/10.1590/S0100-204X2007000100010

Rosa, V. G. C., Moreira, M. A., Rudorff, B. F. T., & Adami, M. (2010). Estimation of coffee productivity based on an agrometeorological spectral model. *Pesquisa Agropecuaria Brasileira*, *45*(12), 1478–1488. https://doi.org/10.1590/S0100-204X2010001200020

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *Progress Report RSC 1978-1*, 112. https://ntrs.nasa.gov/search.jsp?R=19740022555

Santos, M. A. dos, & Camargo, M. B. P. de. (2006). Parameterization of an agrometeorological

model for estimating coffee yield in the conditions of the State of São Paulo. *Bragantia*, v. 65, n. 1, p. 173-183. http://dx.doi.org/10.1590/S0006-87052006000100022

Schattan, S. (1964). Research for an objective method to forecast coffee production. *Agricultura em São Paulo, São Paulo*, *11*(3–4), 1–43.

Setiyono, T. D., Quicho, E. D., Gatti, L., Campos-Taberner, M., Busetto, L., Collivignarelli, F., et al. (2018). Spatial rice yield estimation based on MODIS and Sentinel-1 SAR data and ORYZA crop growth model. *Remote Sensing*, *10*(2), 293. https://doi.org/10.3390/rs10020293

Silva, M. G., Pozza, E. A., & Vasco, G. B. (2019). Geostatistical analysis of coffee leaf rust in irrigated crops and its relation to plant nutrition and soil fertility, 117–134.

Silva, S. de A., de Souza Lima, J. S., & de Oliveira, R. B. (2011). Agrometeorological model in estimating the productivity of two varieties of Arabica coffee considering spatial variability. *Irriga*, 1–10. https://doi.org/10.15809/irriga.2011v16n1p01

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*, *21*(2), 110–124. https://doi.org/10.1016/j.tplants.2015.10.015

Sugawara, L. M. (2002). Evaluation of agrometeorological model and NOAA/AVHRR images in the monitoring and estimate of soybean productivity in the state of Paraná. *São José dos Campos*, 181. http://bibdigital.sid.inpe.br/rep-/dpi.inpe.br/lise/2002/11.18.18.05

Tan, B., Masek, J. G., Wolfe, R., Gao, F., Huang, C., Vermote, E. F., et al. (2013). Improved forest change detection with terrain illumination corrected Landsat images. *Remote Sensing of Environment*, *136*, 469–483. https://doi.org/10.1016/J.RSE.2013.05.013

Ustin, S. L., Gitelson, A. A., Jacquemoud, S., Schaepman, M., Asner, G. P., Gamon, J. A., & Zarco-Tejada, P. (2009). Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment*, *113*, S67–S77. https://doi.org/10.1016/j.rse.2008.10.019

Varzea, V. M. P., Rodrigues, C. J., & Lewis, B. G. (2002). Distinguishing characteristics and vegetative compatibility of Colletotrichum kahawe in comparison with other related species from coffee. *Plant Pathology*, *51*(2), 202–207. https://doi.org/10.1046/j.1365-3059.2002.00622.x

Vegro, C. L. R., & de Almeida, L. F. (2019). Global coffee market: Socio-economic and cultural dynamics. In *Coffee Consumption and Industry Strategies in Brazil: A Volume in the Consumer Science and Strategic Marketing Series* (pp. 3–19). Elsevier. https://doi.org/10.1016/B978-0-12-814721-4.00001-9

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of*

Environment, *185*, 46–56. https://doi.org/10.1016/j.rse.2016.04.008

Victorino, E. C., de Carvalho, L. G., & Ferreira, D. F. (2016). Agrometeorological modeling for coffee productivity forecast in the south region of Minas Gerais state. *Coffee Science*, *11*(2), 211–220. https://doi.org/10.25186/cs.v11i2.1049

Wit, C. T. (1965). *Photosynthesis of leaf canopies*. *Agricultural Research Reports*. Pudoc. https://doi.org/10.2172/4289474

COMPUTER CODE AVAILABILITY

Name of code: Yield coffee by Random Forest

- Developers: Jonathan da Rocha Miranda
- Contact details – Department of Agricultural Engineering, Federal University of Lavras, University Campus, PO Box 3037, ZIP Code: 37200-000, Lavras, Minas Gerais, Brazil, email: jhonerocha@estudante.ufla.br
- Year first available: 2020
- Hardware required: Yield coffee by Random Forest was run on a computer with 4 cores (2.4 GHz each) and 4 GB
- Software required: Yield coffee by Random Forest was interpreted with Spyder and needs Pandas, NumPy, Seaborn and Matplotlib packages
- Program language: the code is written in Python 3.6
- Program size: 463 KB
- Details on how to access the source code: the source files of the Yield coffee by Random Forest can be downloaded from GitHub: https://github.com/jonathanrocha71/Yield_coffe_Random_Forest.git