BRUNO DE OLIVEIRA SCHNEIDER

COFFEE MAPPING BY REMOTE SENSING AND MACHINE LEARNING

LAVRAS – MG
2023

BRUNO DE OLIVEIRA SCHNEIDER

COFFEE MAPPING BY REMOTE SENSING AND MACHINE LEARNING

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

Prof. Dr. Marcelo de Carvalho Alves
Orientador

LAVRAS – MG
2023

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

Schneider, Bruno de Oliveira
Coffee mapping by remote sensing and machine learning / Bruno de Oliveira Schneider. – Lavras : UFLA, 2023.
43 p. : il.

Orientador: Prof. Dr. Marcelo de Carvalho Alves.

Tese(doutorado)–Universidade Federal de Lavras, 2023.
Bibliografia.

1. Sensoriamento Remoto. 2. Mapeamento de Café. 3. Aprendizado de Máquina. I. Alves, Marcelo de Carvalho. II. Título.

BRUNO DE OLIVEIRA SCHNEIDER

COFFEE MAPPING BY REMOTE SENSING AND MACHINE LEARNING

IDENTIFICAÇÃO DE CULTIVO DE CAFÉ POR SENSORIAMENTO REMOTO E APRENDIZAGEM DE MÁQUINA

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

APROVADA em 14 de Agosto de 2023.

Prof. Dr. Fortunato Silva de Menezes	DFI/UFLA
Prof. Dr. José Sérgio de Araújo	IFSULDEMINAS
Dr. Gladyston Rodrigues Carvalho	EPAMIG
Dr. Luciano Teixeira de Oliveira	SEMAD/Prefeitura M. de Contagem

gov.br Documento assinado digitalmente
MARCELO DE CARVALHO ALVES
Data: 09/11/2023 17:04:18-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo de Carvalho Alves
Orientador

LAVRAS – MG
2023

Dedico este trabalho aos meus pais Sérgio (in memoriam) e Marilena, que me incentivaram e me apoiaram ao longo de tantos anos para que fosse possível chegar até aqui.

AGRADECIMENTOS

Agradeço à minha família, Andréia e Mariana, pela paciência e compreensão. Agradeço ao meu orientador, Marcelo, pela disposição em orientar alguém de outra área e sem dedicação integral. Agradeço ao PPG em Engenharia Agrícola e à UFLA pela oportunidade de aperfeiçoamento.

RESUMO

O cultivo de café é um elemento importante na economia brasileira, produzindo receitas de exportação, empregos e movimentando a economia local. O Brasil é o maior produtor e exportador de café do mundo e o seu plantio é um elemento importante na nossa identidade cultural. Conhecer com precisão as regiões de cultivo de café permite avaliar melhor o balanço entre a oferta e demanda do produto, facilita monitorar os problemas associados ao cultivo, auxiliando na tomada de medidas preventivas ou determinação de políticas públicas para direcionar essa atividade. O mapeamento, feito de forma tão automática quanto possível, ajudaria a manter dados atualizados sobre as regiões de cultivo, em grandes extensões de terra. Por outro lado, o mapeamento automático desta cultura enfrenta diversos desafios relacionados com a variedade de características de produção de café em diferentes locais. As variações vão desde o uso de diferentes espécies e variedades de plantas, como diferenças visuais relativas à idade das plantas, cultivo em consórcio com outras culturas e técnicas de cultivo e manejo, como o plantio sombreado. Este trabalho investiga o mapeamento de cultivo de café, na forma de monocultura exposta, no município de Lavras, MG, usando imagens do satélite Sentinel-2 (MSI) e o algoritmo de classificação Random Forest. O Random Forest é um algoritmo de aprendizado de máquina e portanto "aprende" a fazer a classificação por meio de exemplos, o que requer uma pequena classificação manual a fim de gerar exemplos que pretendam abranger os vários casos possíveis de classificação. Produzir uma amostragem adequada à criação de uma classificação cria problemas práticos diversos, cujo impacto na classificação ainda precisa ser melhor estudado. Neste trabalho, observamos que alguns aspectos práticos tem efeitos bem mais significativos que outros. Foram realizados testes de classificação, mostrando que a escolha de exemplos de classificação acaba produzindo efeitos mais significativos do que a escolha das bandas eletromagnéticas amostradas pelo satélite. A inclusão de ruído (sombras, falhas no plantio, carreadores) nas amostras de plantio não impediu a classificação adequada. Foi também desenvolvida uma técnica para eliminar o ruído comum nas classificações por pixel, produzindo áreas mais contínuas de classificação, que são mais apropriadas para demarcação geométrica. A análise de acurácia se concentrou na classificação de uma região distinta do treinamento, uma característica pouco comum em trabalhos anteriores, mas que é importante para a viabilidade prática da classificação, uma vez que não é viável produzir uma classificação manual numa grande região a fim de fazer o treinamento do classificador. Foram obtidos resultados de classificação com acurácia de até 94,4%, com Kappa de 0,761, para classificação em região distinta da de treinamento. O sistema de classificação foi todo implementado com software livre, usando dados de satélite que estão publicamente disponíveis, usando a linguagem R e suas bibliotecas, incluindo uma implementação de Random Forest da biblioteca `ranger`.

Palavras-chave: Mapeamento. Café. Uso da Terra. Sensoriamento Remoto. Sentinel-2. Random Forest. Coerência Espacial.

ABSTRACT

Coffee cultivation is an important element in the Brazilian economy, producing export revenues, jobs and driving the local economy. Brazil is the largest producer and exporter of coffee in the world and its cultivation is an important element in our cultural identity. Knowing precise locations of coffee growing allows a better assessment of the balance between supply and demand for the product, makes it easier to monitor problems associated with cultivation, helping to take preventive measures or determine public policies to guide this activity. Mapping, done as automatically as possible, would help keep up-to-date data on growing regions across large tracts of land. On the other hand, automatic mapping of this particular crop faces several challenges related to the variety of coffee growth systems in different locations. Variations range from the use of different species and varieties of plants, such as visual differences related to the age of the plants, intercropping with other crops, and cultivation and management techniques. This work investigates the mapping of coffee crops, in the form of sun exposed monoculture, in the municipality of Lavras, MG, using images from the Sentinel-2 MSI satellite and the Random Forest classification algorithm. Random Forest is a machine learning algorithm and therefore "learns" to classify through examples, which requires a little manual classification in order to generate examples that intend to cover the various possible cases of classification. Producing adequate sampling to create a classification creates several practical problems, whose impact on classification still needs to be better studied. In this work, we observed that some practical aspects have much more significant effects than others. Classification tests were carried out, showing that the choice of classification examples ends up producing more significant effects than the choice of electromagnetic bands sampled by the satellite. The inclusion of noise (shadows, planting failures, road tracks) in the crop samples did not lead to a bad classification. A technique was also developed to eliminate common noise in pixel-based classifications, producing more continuous areas of classification, more suitable for geometric demarcation. The accuracy analysis focused on the classification on an area distinct from the training region, an uncommon feature in previous works, but which is important for the practical feasibility of the classification, since it is not feasible to produce a manual classification in a large region to be used. in order to train the classifier. Classification results were obtained with accuracy of up to 94.4%, with Kappa of 0.761, for classification in a region other than the training one. The classification system was all implemented with free software, using satellite data that are publicly available, using the R language and its libraries, including a Random Forest implementation of the `ranger` library.

Keywords: Crop Mapping. Coffee. Land Use. Remote Sensing. Sentinel-2. Random Forest. Spatial Coherence.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

PART ONE

This work is presented in two parts. The first contains a general introduction, the work goals, the hypothesis that motivates this research, the original contributions produced, a quick review of previous work on coffee mapping and finally the conclusions gathered from it. The second part is an scientific article for peer-review magazines for a quick and broader divulgation; it presents in detail the data set, methods and results obtained. The article was reformatted according to UFLA's style standards for thesis documents. Bibliographical references from both parts are at the end, also following UFLA's standards. A R script capable of mapping coffee is included as an appendix for reproducibility purposes.

1 INTRODUCTION

Coffee is an economically important commodity in Brazil, with significant influence on the economy and the occupation of Brazilian land. It is important to know the dynamics of land use for coffee cultivation in order to learn more about the productivity of the sector and how the sector influences different events of land use such as deforestation. Knowing the impact of coffee cultivation involves knowing the areas and locations of this cultivation, which is a complex task in such a large territory. Land use mapping, using satellite images, can significantly contribute to this task, but there are still many difficulties to be overcome in this area. As technologies for automatic mapping evolve, precise and up-to-date data about coffee cultivation may help develop public policies that help economic growth as well as environmental protection, much as it is already been done with deforestation mapping (FINER et al., 2018).

Satellite images are subject to difficulties such as atmospheric interference, especially cloud occlusion, variations in solar radiation and differences in sensor recordings that are used in different satellites. There is also possible occlusion caused by relief features, that is, portions of the Earth's surface hidden by relief at the satellite's observation point. Many kinds of classification analyses can be carried out based on vegetation cover data, in particular we are interested in identifying the type of vegetation, and in the case of crops, the identification of the cultivated species. There are already works for the classification of several species, such as corn (AVCI; SUNAR, 2015), soy (ZHONG et al., 2016), cotton (XUN et al., 2021), sugar cane (SINGH; PATEL; DANODIA, 2020), brachiaria and coffee (MOREIRA et al., 2010). The identification of crops by satellite data allows the mapping of large regions in a short time, and

the mapping of cultures favors the understanding of how the planted areas interact with other types of soil occupation, favors the understanding of the dynamics of agriculture in a region. Knowing the land use in a region helps to identify environmental problems such as deforestation and the consumption of agricultural defensives, allows planning for harvests and allows anticipating the needs of agricultural activity, eventually favoring the development of adequate public policies.

Coffee is an important crop in the region of Lavras, MG, where UFLA is located. The state of Minas Gerais accounts for more than 50% of the national coffee crop. Being able to produce and keep updated an adequate mapping of coffee cultivation in the state of Minas Gerais is therefore important in the economic and environmental context of the region. There are vegetation cover classification initiatives such as MAPBIOMAS[1] which contains data updated until 2019 and does not classify coffee plantations and the Portal do Café de Minas initiative[2] which aimed to identify coffee plantations and was closed in 2018. Identifying coffee plantations is still challenging and there is a lack of accurate, up-to-date, publicly available data on coffee regions. Although coffee classification works already exist, it is not always feasible to try to reproduce existing works. Obstacle examples to use previously described work include: use of proprietary data at a significant cost, regional interference from cloudiness, terrain relief, intercropping of cultures, differences between development stages (phenology), growth stages (age of plants), size of properties and lack of presentation of software used (HUNT et al., 2020).

The above mentioned difficulties affect spectral signature, local image patterns and surface patterns, therefore classification models should be tunned at their particular goal regions, even though the methods are the same. This work intends to carry out the analysis for regions near the municipality of Lavras, MG, Brazil, given the importance of coffee growing in the region and the difficulty for creating ground truth maps without funding.

Data and tools needed for mapping can be expensive to acquire. This work, like so many other at public universities in Brazil, had no funding, so only free software and publicly available data were used, except for coffee plantation polygons, kindly provided by EPAMIG, generated years before by joint EPAMIG/EMATER mapping projects that received public funding. However, because this data was outdated and had several precision problems, coffee plantation polygons were fully reviewed and corrected for the working regions in this research.

[1] https://mapbiomas.org/

[2] http://portaldocafedaminas.emater.mg.gov.br/

This work is an investigation of how accurate can coffee plantations be identified using freely available data and tools, with the following specific goals in mind:

- try the Random Forest (BREIMAN, 2001) algorithm to deal with kind of error generated by demarcating coffee plantations on satellite images, instead of choosing prime samples;
- see how well extrapolation of a classification model created in one region works on other regions;
- investigate how useful are the multi spectral bands available on Sentinel-2 data for coffee classification;
- mitigate expected pixel-based classification errors using some sort of spatial coherence mechanism.

2 HYPOTHESIS

With so many sources of error for coffee classification, tight scope published works are the norm. We investigated if is it viable to detect coffee plantations from publicly available data (Sentinel-2 MSI), using publicly available software (R, Random Forest), using a simple training approach, that is to extract all data available from a manual classification area. Despite its many challenges, we hypothesize that coffee mapping is viable for large areas, using publicly available data and software, on ordinary computers (as opposed to high end computers), with good accuracy, by combining already tried strategies of model tuning, using pixel-based classification.

3 THEORETICAL BACKGROUND

Coffee mapping is evolving since around the year 2000 when satellite images started being used for coffee mapping (ALVES; RESENDE; ANDRADE, 2000). Since then, several new technology advances have provided increasing success in such task.

There are many approaches to coffee mapping using remote sensing data (HUNT et al., 2020):

- Pixel-based, where each pixel[3] is classified separately. It works well with low and mid resolutions but has little spatial coherence and produces mappings in which false positives and false negatives are scattered across the image, in an effect that may be called salt-and-pepper error.

- Sub-pixel, where each pixel is considered a mixture of classes of the classes being identified, in order to overcome limitations of the spatial resolution.

- Texture-based, where the pixels are classified using neighborhood information, where patterns either on the surface (SAR[4] images) or the visual pattern (high-resolution optical images) are used in the classification.

- Object-based, where pixels are first grouped into sets through some properties, and then classified as a set.

- Fusion and hybrid approaches exist, combining different kinds of data (spectral, SAR, relief) or combining the above approaches.

Because there are different types of coffee growing systems, coffee mapping should be capable of classifying more than one coffee class as done in previous work (MASKELL et al., 2021; ESCOBAR-LÓPEZ et al., 2022).

As newer satellite sensing instruments become capable of sensing more spectral bands, choosing bands for use in a classification model impacts its accuracy. Choosing bands is usually done by testing correlation or Principal Component Analysis (PCA). However such results are bound to change as a factor of the sampled data, moving classification to a different area, with different classification samples, or adding new training samples for model creation would make changes in how useful any band is in the model. Spectral bands are often associated with a particular usefulness, for instance, Sentinel-2's band 10 (1363.5 nm) is known for cirrus cloud detection because it gets reflected by water in high altitudes and absorbed by water closer to the ground. However, given the absence of cirrus clouds, it is unclear whether this band could help differentiate coffee plants from some other vegetation. We decide to use Random Forest's own variable importance ranking to choose which bands to use. Results differ slightly depending

[3] A pixel is the smallest data in a set that define some visual information, such as an image. In remote sensing, a pixel is associated with the smallest sensing area and contains information from all spectral bands that the sensing instrument is able to capture.

[4] Synthetic Aperture Radar (or just radar for short)

on the training data, but for the training data used for the results shown later, the only band removed from the data was band 10. As an example for contrast, Escobar-López et al. also removed bands 1 (440nm) and 12 (2200nm) from their data (ESCOBAR-LÓPEZ et al., 2022).

Other strategies to cope with the difficulties for coffee classification, include using multi-temporal data (MOREIRA; BARROS; RUDORFF, 2008; BERNARDES et al., 2012; ORTEGA-HUERTA et al., 2012), so that phenological changes trough different seasons are captured. Cloud coverage is also mitigated by fusing data for different days, specially when SAR data is also present (MASKELL et al., 2021). Ad hoc strategies also were found, for instance, Andrade et al. mapped water lines and made a 50m buffer around them to mask coffee identification, because such areas were protected by law in the studied area, and forest has similar spectral properties as coffee crops (ANDRADE et al., 2011).

Texture (SILVA et al., 2009; LELONG; THONG-CHANE, 2003) or Object-based (VIEIRA et al., 2007; SANTOS et al., 2012) classifications benefit from creating sets of spatially close pixels with the same classification. However they involve significantly more computation and more complex tuning of the classification model creation. Poor choice of features can lead to whole sets of pixels being misclassified. Comparison of pixel-based versus texture or object-based tend to favor set classification (GAERTNER et al., 2017; ARIAS, 2007). The term texture is also used with two different meanings, it could be statistical attributes such as standard deviation, contrast, entropy, etc. (MARUJO et al., 2017) which in turn may used for image segmentation into objects (CHAKRABORTY; SEN; HAZRA, 2009), but it also could mean frequency attributes that measure directionality (TSAI; CHEN, 2017).

The algorithms used for classification also vary greatly, including Support Vector Machine, Multi Layer Perceptron (BOELL, 2016), Random Forest and others. There are some comparisons already made (SOUZA et al., 2016) but accuracy results vary significantly between studies, especially because there are many ways of measuring accuracy. There has been published works on how sampling affects Machine Learning classification (RAMEZAN; WARNER; MAXWELL, 2019) but different classification algorithms are affected differently by sampling.

Coffee mapping has evolved from RGB thresholding in the years 2000 to multi-spectral machine learning in two decades, but there is still room for much improvement. A few works, previously cited in this section, are shown in Table 1 to illustrate both evolution in the state of the and comparison to this work.

Table 1 – Instances in timeline of coffee mapping

Year	1st author	Remarks
2000	Alves	RGB threshold classification.
2007	Vieira	Used proprietary software, concludes that remote sensing is useful for mapping.
2008	Moreira	Used multi-temporal images and maximum likelihood estimation. Proprietary software.
2011	Andrade	Used Artificial Neural Network and proprietary software. Used terrain relief to rule out coffee from nearby water streams because those areas would be protected.
2016	Boell	Compared Multi-layer Perceptron and Support Vector Machine on proprietary software. Used general terrain cover classes including coffee. Not much detail about how the manual classification was done.
2016	Souza	Compared variables and algorithms for classification. Used proprietary data and software. Used general classes. Concluded that coffee is more often confused with forest and that spectral variables are more correlated than textural ones.
2017	Marujo	Used object segmentation. Effort on separating distinct coffee classes. Highlights texture advantages. Similar results on all coffee classes..
2017	Tsai	Proposed two frequency indices for classification.
2017	Baeta	Deep Learning (Convolution Neural Networks). Processing on GPU. Used classifiers on different size windows.
2021	Maskell	Used Sentinel-1 & 2 data fusion, multi-temporal, Random Forest classifier.
2022	Lopez	Used Sentinel-2 plus radar with Random Forest, multi-temporal.

4 METHODOLOGY

Any machine-learning based classification and mapping must consider choosing samples for the training phase. Training is the phase where the classification algorithm chooses rules and their parameters that will allow classification. Once training is done. The classification algorithm is configured and is therefore called a classifier. The classifier will then receive new input data, producing a classification output for each input received.

Training data is often chosen from unequivocal data. A classifier used for coffee classes would then receive data samples that are good instances of those coffee classes. But because coffee crops may have very different characteristics, creating many coffee classes that confines that variation into homogeneous classes is often the logical approach. However, obtaining information about the location of coffee crops is difficult enough. Generating that information for each coffee class might be unfeasible because a single coffee plot may have a subarea with unhealthy or underdeveloped plants. A coffee crop may contain trees, their shadows, and bare

soil tracks which are unlikely to be separated from the coffee plants in the training data. Crop borders are likely to have different characteristics from the internal areas as pixels in the remote sensed data would be a mixture of characteristics from coffee plants and bare soil pathways or whatever is at the border.

We have generated training data by manually drawing coffee polygons over Sentinel-2 data for each coffee class. This manual classification was done by expert interpretation using high resolution, freely available data that was only used at this phase. Data sources were CBERS-4A satellite, Google Maps imagery, and old vector mapping data from previous mapping projects at Brazilian research institutes. The variety of sources, which span for a time period of at least a year, produced a high confidence in the manual classification.

Three distinct regions were fully classified manually. They are shown in Figure 2 in the article at Part Two. All pixels from *training region 1* were used as input for training. Only coffee pixels from regions two and three were used as input, as way of increasing the number of coffee samples in the training.

Choosing training data if a tricky situation because if only perfect samples of coffee are used in the training, the resulting mapping would shrink coffee areas and produce holes in coffee patches. On the other hand, samples from trees, tree shadows and roads, identified as coffee for the training increase the likelihood of false positives during the classification phase.

In this work, the manual classification consisted of drawing polygons over coffee crops, without regard for noise inside or on the borders. Only big features such as a large building surrounded by crops were marked. Crop polygons then must be converted into raster data (pixels) before training. Only pixels completely covered by the polygons were considered for their respective coffee classes, as is done by the rasterization algorithm in the `terra` library for R language. This reduces the areas a little, but does not impose a strict classification method. When the coffee polygons are created, it is expected that some pixels on the borders are noisy, that is, are poor instances for the coffee class. But because noise would also be expected inside the polygons for unhealthy plants, tracks and trees, we just assume the training must deal with noise.

The Random Forest algorithm is known for being robust to noisy data, as classification is done independently several times and the final output derives from majority. Good results are obtained as long as the noisy samples are in small proportion compared to the good samples.

In order to provide a large amount of good samples, all samples from coffee classes available at training were used. The training is simplified because manual classification is not very strict and also there is no need for a strategy for choosing training samples. During experimentation, it was noticed that a large number of samples were not a significant computing obstacle.

The choice of spectral bands to use for training and mapping is also a common concern in similar work. We assumed that even though some bands would have little correlation to the classification, any information might be helpful. Using all spectral bands of Sentinel-2 did not prove to be a computing obstacle either. As the Random Forest algorithm has its own importance ranking for variables and the `ranger` library could compute that rank automatically during training. We only tested how accuracy varied from removal of the lower rank band. We noticed that a minor enhancement was produced from removal of the less important band, and that improvement quickly turns into accuracy loss as lower importance bands keep being removed.

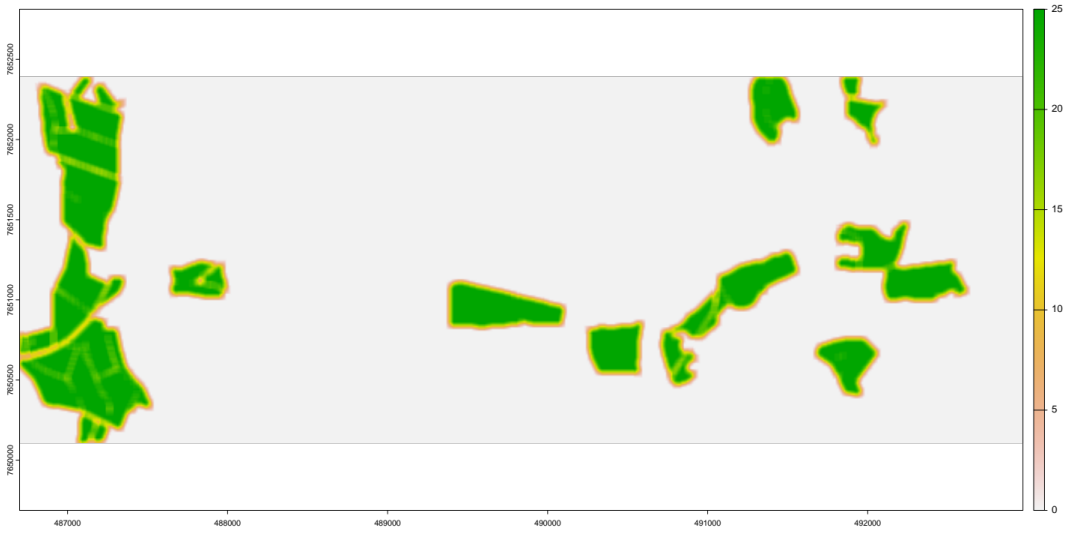
In this work, instead of using statistical samples and statistical accuracy indices, we decided to compute every pixel in the working region. Randon Forest resistance to over-fitting was put to the test, by using every pixel in the *training region 1* for model training, as well as every coffee pixel in the other training regions. Every pixel in the classification region was subjected to classification and used to compute output accuracy. Even so, how to measure accuracy could still be subject to debate. For instance, we had two coffee classes but if a sample from one coffee class is classified as the other coffee class, is it a misclassification? We opted to consider it a misclassification, honoring the class separation.

Because pixel-based classification cannot use information from the general area where the pixel is in, texture-based methods have been proposed. We experimented on a novel way to use the area information. Ambiguous values for spectral data would be better classified if spatial coherence was used, that is, a pixel is more likely to be a member of a coffee crop if its neighbors are also members of the coffee crop.

Given any area classification, this information could be represented as the number of neighbor pixels that have been classified as member of a given coffee class. We call this number the neighborhood information. Figure 1 show neighborhood information for the *training region*, using a 5x5 neighborhood matrix. Considering such area, any pixel's neighborhood information

would be an integer number ranging from zero to 25, which indicates how many neighbors (including the pixel itself) belong to a specific coffee class.

Figure 1 – Neighborhood values for the training region.



Neighborhood information is computed after some classication, by aplying a focal operation on a classification map. A focal operation is a raster operation where each pixel gets a new value as a function of the value of all pixels inside a window surrounding it. In this case, all 25 pixels of a 5x5 windows whose center is the pixel being computed have their values added, then the sum is assigned as a new value for the pixel. The inicial raster data is a matrix where each pixel that belongs to a specific coffee class is assigned a value of one and all other pixels are assigned a value of zero.

However, a classification is needed for those initial raster values, while neighborhood information is needed for classification. We circumvented this problem by doing a classification in two or more steps, as shown in the article (Part Two), on the methodology overview (Figure 5) and in section 4.2. At each iteration, classification gets a little better, although by an ever smaller amount, so it might not be worth the processing time to go further than the second or third iteration. Images showing the classification improvement are also in Part Two (Figures 7 and 8).

Therefore, our pixel-based classification is enhanced by neighborhood information, that gives it a spatial coherence and avoids the salt-and-pepper effect on classification results.

5 ORIGINAL CONTRIBUTIONS

Previous coffee mapping work, to the best of our knowledge, has not been done using Sentinel-2 (MSI) full multi-spectral data. While we found pixel-based, random forest classification using Sentinel-2 data, details about how bands were chosen are unclear. We used Random Forest's own internal variable importance ranking for choosing spectral bands.

Mapping coffee using Random Forest has not been published using an area that is distinct from the training area, which is an important practical obstacle for mapping large areas.

We proposed a novel technique for using spatial coherence as a way to improve mapping accuracy, one that is independent of the crop being identified, and that does not rely on visual patterns created by the trees because, for coffee, tree spacing has significant variation and terrain relief impact the pattern formed (BAETA et al., 2017).

Details of software (R language and its libraries), working region, the extrapolation of the classification model to a nearby but disjunct region also set this work apart from previous coffee classification in the literature.

6 CONCLUSIONS

Coffee identification using Random Forest classification has shown to be able to deal with high levels of noise produced during manual classification that generates the training data. Nevertheless, the training data has shown to be the most important factor for good accuracy. We propose that instead of worrying about shadows and clear spots inside plantations, training efforts should concentrate on finding relatively enough distinct coffee samples compared to the non-coffee samples. Because contiguous areas may have few coffee pixels, our solution to increasing the ratio of coffee samples was to add only coffee samples from other areas.

We made accuracy comparisons when removing bands with lower computed importance during classification. Because an unexpected band ranked high for classification, further work needs to be done in investigating what could be the best band subset to use in training for coffee identification. Using almost all data from multi-spectral MSI instrument in Sentinel-2 satellites enhances classification accuracy. The accuracy gained from removing the least important band (band 10) is only marginal.

Adding vegetation indices to training data also showed a marginal accuracy increase for coffee classification. Further work on experimenting which indices subset would be more useful is recommended.

Neighborhood data is very useful for elimination of salt-and-pepper error. By computing how many neighbors of each coffee class and using that information during training, the classification model will more likely keep the pixels that were classified along with its neighbors on the same class. The resulting classification is more accurate and resembles that of an object-based classification without the need for finding the objects in advance.

REFERENCES

ALVES, H. M. R.; RESENDE, R. J. T. P. de; ANDRADE, H. Utilização do SPRING para avaliação do uso da terra em agroecossistemas cafeeiros da região de São Sebastião do Paraíso-MG. In: **Simpósio de Pesquisa dos Cafés do Brasil**. [s.n.], 2000. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/902679/utilizacao-do-spring-para-avaliacao-do-uso-da-terra-em-agroecossistemas-cafeeiros-da-regiao-de-sao-sebastiao-do-paraiso---mg>.

ANDRADE, L. N. de et al. Redes neurais artificiais (RNA) aplicadas à classificação de áreas cafeeiras na região de Três Pontas-MG. In: **VII Simpósio de Pesquisa dos Cafés do Brasil**. [S.l.: s.n.], 2011.

ARIAS, S. B. **Using Image Analysis and GIS for Coffee Mapping**. McGill University Libraries, 2007. (McGill theses). Disponível em: <https://books.google.com.br/books?id=xcjtoQEACAAJ>.

AVCI, Z. D. U.; SUNAR, F. Process-based image analysis for agricultural mapping: A case study in turkgeldi region, turkey. **Advances in Space Research**, v. 56, n. 8, p. 1635–1644, 2015. ISSN 0273-1177. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0273117715005232>.

BAETA, R. et al. Learning deep features on multiple scales for coffee crop recognition. In: **2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2017. p. 262–268.

BERNARDES, T. et al. Monitoring biennial bearing effect on coffee yield using modis remote sensing imagery. **Remote Sensing**, v. 4, n. 9, p. 2492–2509, 2012. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/4/9/2492>.

BOELL, M. G. **Sistema de classificação de imagens de sensores remotos**. 91 p. Dissertação (Mestrado em Engenharia de Sistemas e Automação) — Universidade Federal de Lavras, Lavras, MG, 2016.

BOURGOIN, C. et al. Assessing the ecological vulnerability of forest landscape to agricultural frontier expansion in the central highlands of vietnam. **International Journal of Applied Earth Observation and Geoinformation**, v. 84, p. 101958, 2020. ISSN 1569-8432. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0303243419307202>.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CHAKRABORTY, D.; SEN, G. K.; HAZRA, S. Texture analysis for very high spatial resolution panchromatic satellite image segmentation. **Asian Journal of Geoinformatics**, v. 9, n. 4, 01 2009.

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 International**, Taylor & Francis, v. 32, n. 7, p. 759–776, 2017.

ESCOBAR-LÓPEZ, A. et al. Identifying coffee agroforestry system types using multitemporal sentinel-2 data and auxiliary information. **Remote Sensing**, v. 14, n. 16, 2022. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/14/16/3847>.

FINER, M. et al. Combating deforestation: From satellite to intervention. **Science**, v. 360, n. 6395, p. 1303–1305, 2018. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aat1203>.

GAERTNER, J. et al. Vegetation classification of coffea on hawaii island using worldview-2 satellite imagery. **Journal of Applied Remote Sensing**, SPIE, v. 11, n. 4, p. 046005, 2017. Disponível em: <https://doi.org/10.1117/1.JRS.11.046005>.

HUNT, D. A. et al. Review of remote sensing methods to map coffee production systems. **Remote Sensing**, v. 12, n. 12, 2020. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/12/12/2041>.

KELLEY, L. C.; PITCHER, L.; BACON, C. Using google earth engine to map complex shade-grown coffee landscapes in northern nicaragua. **Remote Sensing**, v. 10, n. 6, 2018. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/10/6/952>.

LELONG, C. C. D.; THONG-CHANE, A. Application of textural analysis on very high resolution panchromatic images to map coffee orchards in uganda. In: **2003 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2003, Toulouse, France, July 21-15, 2003**. IEEE, 2003. p. 1007–1009. Disponível em: <https://doi.org/10.1109/IGARSS.2003.1293994>.

MARUJO, R. et al. Mapeamento da cultura cafeeira por meio de classificação automática utilizando atributos espectrais, texturais e fator de iluminação. **Coffee Science**, v. 12, p. 17–28, 06 2017.

MASKELL, G. et al. Integration of sentinel optical and radar data for mapping smallholder coffee production systems in vietnam. **Remote Sensing of Environment**, v. 266, p. 112709, 2021. ISSN 0034-4257. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0034425721004296>.

MOREIRA, M. A.; BARROS, M. A.; RUDORFF, B. F. T. Geotecnologias no mapeamento da cultura do café em escala municipal. **Sociedade & Natureza**, Editora da Universidade Federal de Uberlândia - EDUFU, v. 20, n. 1, p. 101–110, Jun 2008. ISSN 1982-4513. Disponível em: <https://doi.org/10.1590/S1982-45132008000100007>.

MOREIRA, M. A. et al. Geotecnologias para mapear lavouras de café nos estados de minas gerais e são paulo. **Engenharia Agrícola**, v. 30, p. 1123–1135, 2010. ISSN 1809-4430.

ORTEGA-HUERTA, M. A. et al. Mapping coffee plantations with landsat imagery: an example from el salvador. **International Journal of Remote Sensing**, Taylor & Francis, v. 33, n. 1, p. 220–242, 2012. Disponível em: <https://doi.org/10.1080/01431161.2011.591442>.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. **Remote Sensing**, 2019.

SANTOS, J. A. dos et al. Multiscale classification of remote sensing images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 50, n. 10, p. 3764–3775, 2012.

SILVA, W. F. et al. Discrimination of agricultural crops in a tropical semi-arid region of brazil based on l-band polarimetric airborne sar data. **ISPRS Journal of Photogrammetry and**

Remote Sensing, v. 64, n. 5, p. 458–463, 2009. ISSN 0924-2716. Theme Issue: Mapping with SAR: Techniques and Applications. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0924271608000932>.

SINGH, R.; PATEL, N.; DANODIA, A. Mapping of sugarcane crop types from multi-date irs-resourcesat satellite data by various classification methods and field-level gps survey. **Remote Sensing Applications: Society and Environment**, v. 19, p. 100340, 2020. ISSN 2352-9385. Disponível em: <https://www.sciencedirect.com/science/article/pii/S235293852030015X>.

SOUZA, C. G. et al. Algoritmos de aprendizagem de máquina e variáveis de sensoriamento remoto para o mapeamento da cafeicultura. **Boletim de Ciências Geodésicas**, Universidade Federal do Paraná, v. 22, n. 4, p. 751–773, Oct 2016. ISSN 1982-2170. Disponível em: <https://doi.org/10.1590/S1982-21702016000400043>.

TSAI, D.-M.; CHEN, W.-L. Coffee plantation area recognition in satellite images using fourier transform. **Computers and Electronics in Agriculture**, v. 135, p. 115–127, 2017. ISSN 0168-1699.

VIEIRA, T. G. C. et al. Geothecnologies in the assessment of land use changes in coffee regions of the state of Minas Gerais in Brazil. **Coffee Science**, v. 2, n. 2, p. 142–149, 2007. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/13717>.

WANG, W. et al. Improving object-based land use/cover classification from medium resolution imagery by markov chain geostatistical post-classification. **Land**, v. 7, n. 1, 2018. ISSN 2073-445X. Disponível em: <https://www.mdpi.com/2073-445X/7/1/31>.

XUN, L. et al. Mapping cotton cultivated area combining remote sensing with a fused representation-based classification algorithm. **Computers and Electronics in Agriculture**, v. 181, p. 105940, 2021. ISSN 0168-1699. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169920331458>.

ZHONG, L. et al. Automated mapping of soybean and corn using phenology. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 119, p. 151–164, 2016. ISSN 0924-2716. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0924271616301071>.

PART TWO

This part is a preliminary version of the article:

Coffee mapping using Sentinel-2 data and spatial coherence

1 Abstract

In this work, we report findings gathered by experiments on mapping coffee plantation using Sentinel-2 data and Random Forest algorithm classification for the region near the city of Lavras, MG, Brazil. We cover extrapolating the classification model to nearby areas for which no samples were used for training, selecting spectral bands and using vegetation indices. We also report a new methodology, for increasing accuracy by spatial coherence that is appropriate to any pixel based mapping of crops. We hope to help shorten the gap between current coffee classification studies and the production of large maps for coffee crops.

2 Introduction

The coffee sector is a significant part in Brazil's economy. Keeping track of coffee plantations and how it changes over time is useful for establishing public policies and understanding how the sector responds to them, similiar to what is been done with deforestation (FINER et al., 2018). Satellite remote sensing data have been used to map coffee plantations and could play a major role in mapping coffee over vast areas and it has been done with reported high accuracy, despite its many challenges. However, accurate, up to date maps are not easy to find, as there seems to be a gap between academic classifications and practical mapping.

Random Forest classification (BREIMAN, 2001) for land use is common in the literature and has been used to map coffee (CHEMURA; MUTANGA, 2017; KELLEY; PITCHER; BACON, 2018; BOURGOIN et al., 2020) under different conditions, most commonly for classifying very distinct classes such as water bodies, urban areas, bare soil and few distinct crops. Nonetheless, since there are different coffee growing system with distinct spectral signatures, coffe is often more than one class, such as in the work of Using distinct classes from a spectral point of view, tends to produce good accuracy. However, since useful classes are very subjective, we wanted wanted to experiment how well Random Forest classification would be for detecting a single crop: coffee.

3 Study area and Data

The experiments reported here used data from Sentinel-2 MSI instrument, tile 23KMS, recorded on June 25, 2022, which is almost free of clouds and areas within that tile which were used were all free of clouds. The data is near the city of Lavras, MG, Brazil, and contains many coffee plantations. Coffee mapping was made by classification using random forests algorithm.

We obtained polygons that demarcate coffee plantations from previous mappings from EMATER/EPAMIG projects. This data used to be available from *Geoportal do Café*[1].

Other satellite images were used for manual classification and processing was done using `R` language version 4.2.2, `Terra` library version 1.7.3 (for geographic raster and vector processing) and `ranger` library version 0.14.1 (random forests implementation).

Since polygons were 3 years old, they were manually verified and edited by using high resolution images from CBERS-4A satellite (pan sharpened 2m resolution) and from Google Maps (resolution not disclosed, but higher then CBERS4A in the region). Four manually verified regions were used for the data reported here, with extents in WGS 84/UTM zone 23S being:

- training region 1: 486700 to 492624 x, 7650110 to 7652390 y (1430.7 ha);
- training region 2: 498534 to 500530 x, 7646218 to 7647919 y (363.8 ha);
- training region 3: 500298 to 502017 x, 7657617 to 7659499 y (333.2 ha);
- classification region: 504138 to 509636 x, 7648014 to 7650800 y (1587.8 ha).

These regions were verified for new plantations, coffee plantations that were put to different use and also coffee plantations polygons were also verified for boundaries, including precision errors from the previous mapping. These regions are around the urban area of Lavras (Figure 2). Some coffee plantations were difficult to verify over satellite images because the plants were too small or had very few leafs. We knew these plantations would share little characteristics with other areas, so they were marked as a different class of coffee (Figure 3). 11 months later, these classes were confirmed to be coffee plantations using new satellite images. When mapping coffee, is usually useful to have classification of more than one class, because properties change a lot over the year and different kinds of coffee systems are likely to appear on classification (ESCOBAR-LÓPEZ et al., 2022).

Figure 2 – Study areas in the city of Lavras shown over the Sentinel-2 RGB data used.

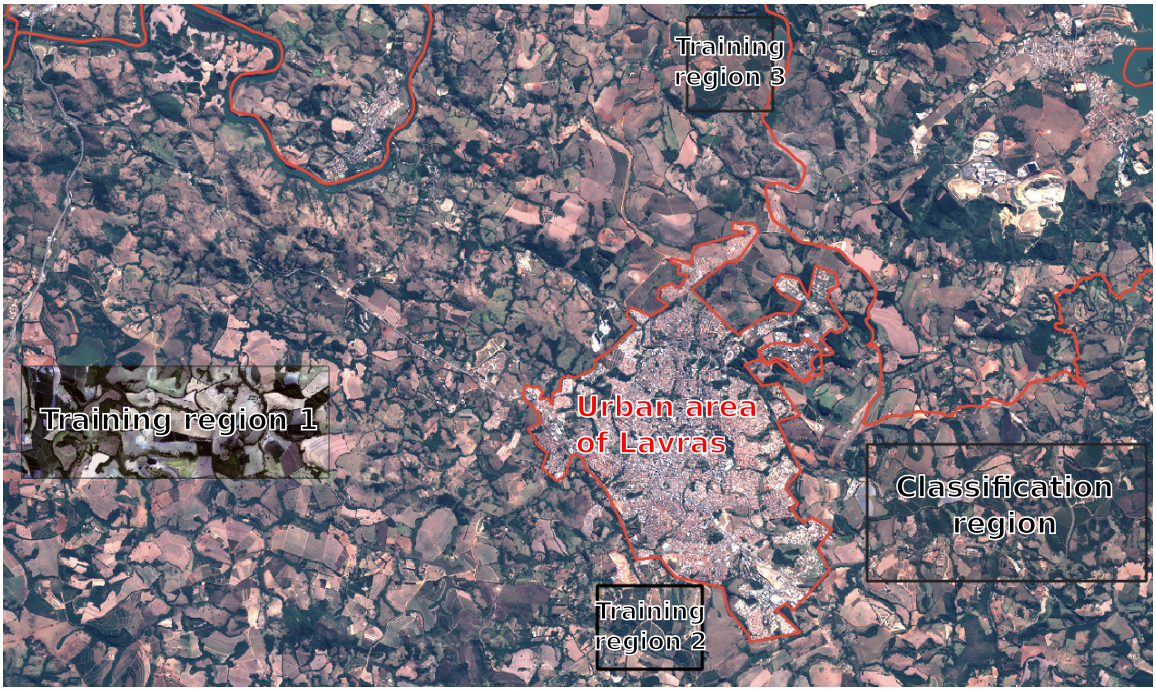
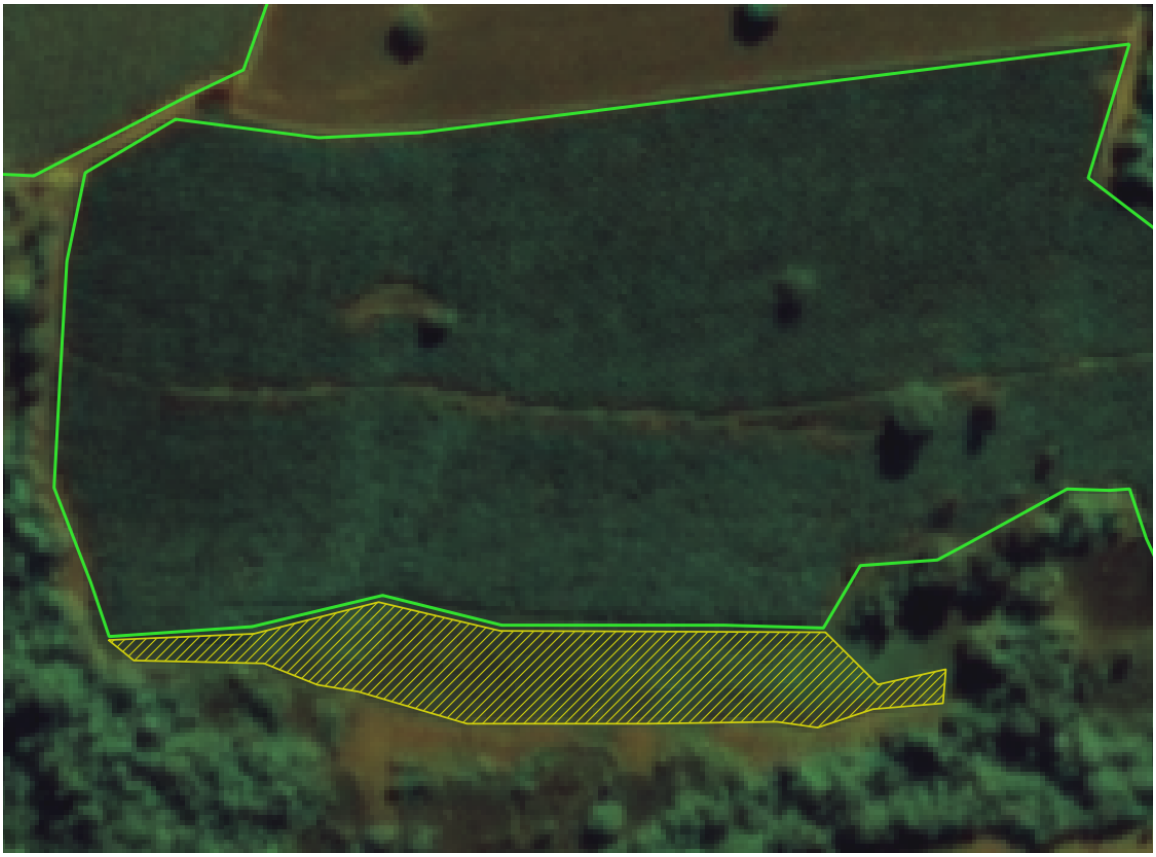


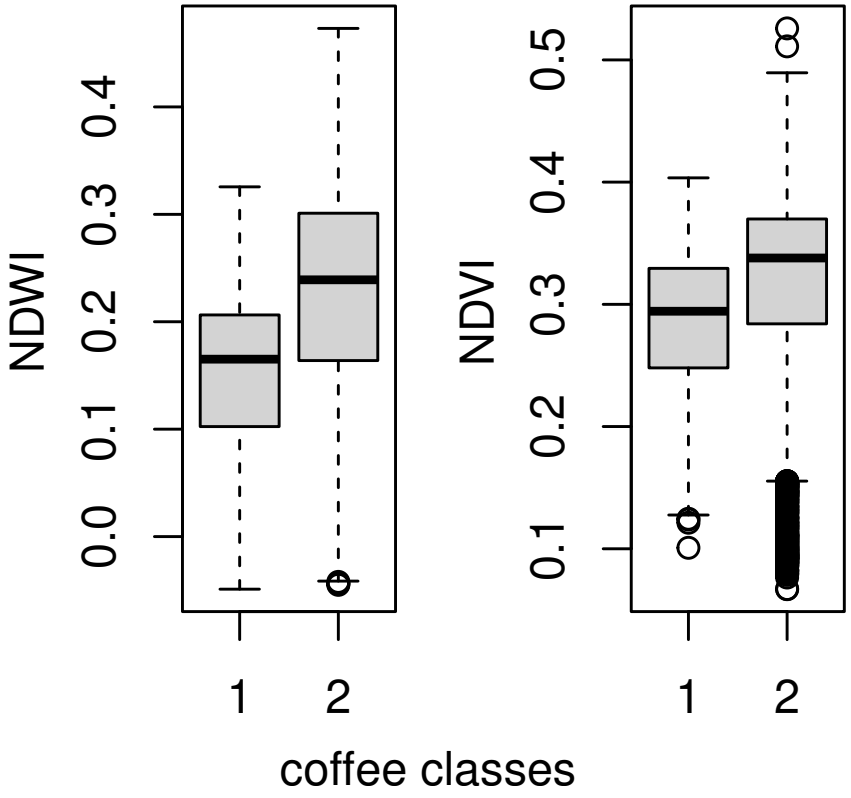
Figure 3 – CBERS-4A pan sharpened image (June 21, 2022) showing coffee plantation polygons of both coffee classes: class 1 (hashed yellow) where coffee canopies were not observable in manual classification; class 2 (green) where canopies are visible.



Coffee polygons contain noise such as big trees inside the plantation, their shadows and farm tracks. They were left as is, for a more practical experiment on coffee mapping than using only pure pixels samples.

We tested NDVI and NDWI vegetation indexes on both coffee classes to confirm they would be distinct in the data (Figure 4).

Figure 4 – Vegetation indexes for the two coffee classes



4 Methodology

We created Random Forest classification models using different sets of bands, including the creation of vegetation indices and our new neighborhood data. Classification was done in the classification region except for confirmation of the *out of bag error* (OOB error) which is an error measurement for the Random Forest algorithm. Accuracy measurements in this work is always producer accuracy.

Due to the cost of creating a manual classification, needed for training and for computing the accuracy of results, small regions of the multi spectral data from Sentinel-2 were used. Using

[1] https://portaldocafedeminas.emater.mg.gov.br/

a localized subset of data of training is considered a valid, effective strategy for classification of geospatial data using machine learning (RAMEZAN; WARNER; MAXWELL, 2019). The training regions 2 and 3 were introduced later on our experiments as a way to provide new, diverse samples and also change proportion of samples of coffee/non coffee classes used for training.

4.1 Traditional methodologies to improve classification

The *training region 1* was used to train a model for classification. All pixels inside the polygons were used as coffee plantation samples (2 classes of coffee), all other pixels were used as "not coffee" class. All 13 Sentinel-2 MSI spectral bands were used in the training. The `ranger` library reported an *OOB error* of 0.0311. To verify if this error measurement would be consistent with overall accuracy, nine tenths of the pixels in the training region 1 were used for training, while the remaining one tenth were used for classification, resulting in an accuracy of 96.9%, compatible with the *OOB error* computed.

Using all pixels from the *training region 1* for training, followed by classification of all pixels in the *classification region* produced an overall accuracy of 87.34%, which was unexpected as both regions are in the same image and therefore have little differences. Trying to improve accuracy, we tested the following three known strategies.

Removing less important bands

As the independent variables may be ranked by importance in the Random Forest algorithm, the importance computed by the `ranger` library was used to choose bands for removal from training, the idea is that using less variables the algorithm with more memory to work the most important bands and would also reduce the noise that the training algorithm has to deal with. Band 10 (SWIR/Cirrus 1.375μm) was ranked the less important band in classification. Removing it resulted in a marginal improvement in accuracy, as report in the Results section.

Adding samples of coffee classes to the training

We used all pixels in training region for training, instead of using a small sample as done usually because we wanted to let the classification manage the noise that naturally occurs on crop delimitation for satellite images. Noise includes crop borders, farm tracks inside the crops, shadows from big trees and differences in vigor of the actual coffee trees. Random Forest is known for its robustness related to overfitting, so a classification would benefit from a large number of examples. Also, the non-coffee class is very diverse, because we did not create distinct classes for urban areas, rivers, lakes and other vegetation covers.

Because the number of samples for non coffee class was far greater and both coffee classes, all coffee pixels from training regions 2 and 3 were added to the training data. Increasing the number of samples had the most improvement in accuracy.

Adding vegetation indexes to the training

Vegetation indexes are commonly used for land cover classification, but because they are not independent data, but instead computed from the bands already available for training, there is always a discussion about its benefits. We chose NDVI and NDWI indexes because the first is a well established index and the latter is related to water. Adding both indexes resulted in a marginal improvement in accuracy.

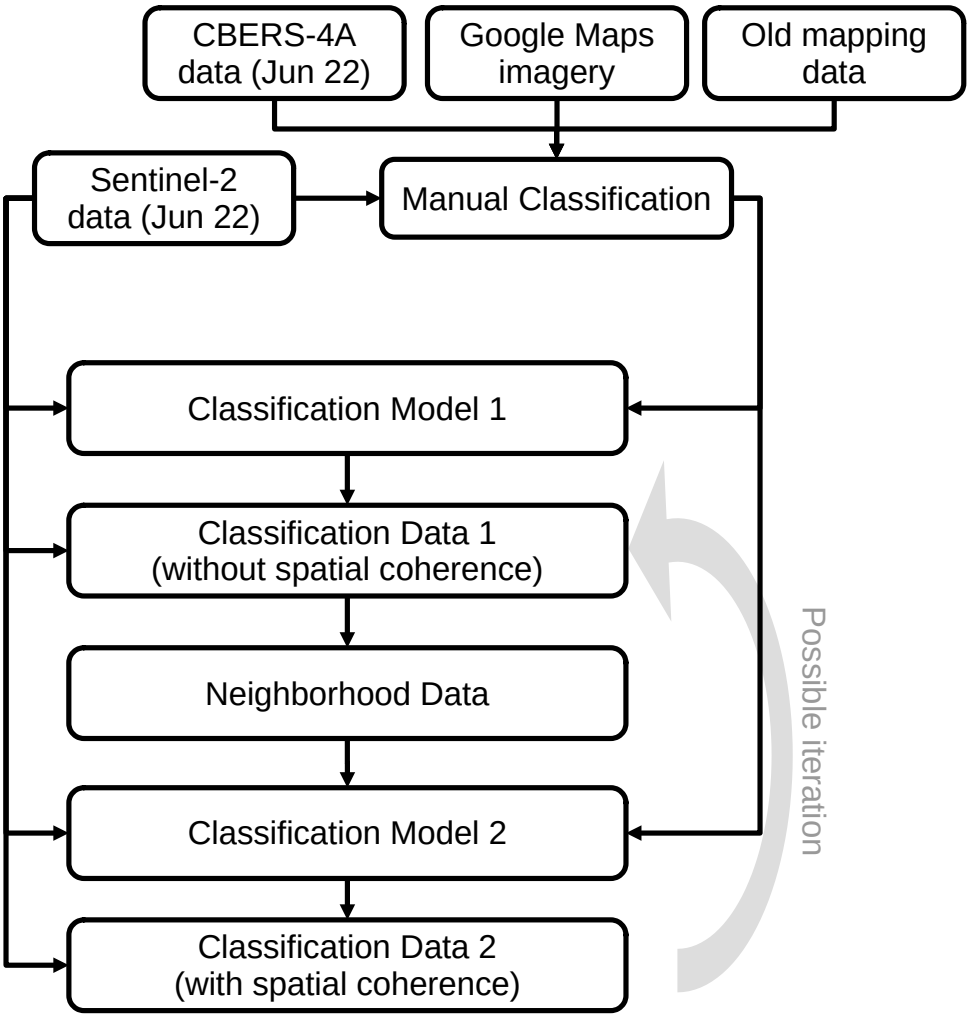
Methodology Overview

The adopted methodology consisted of generating a manual classification from the best data available to us followed by experimenting several tuning procedures that would generate the highest accuracy. Later, neighborhood data was generated from the initial classification and fed back to the model generation, as shown in Figure 5.

4.2 Novel methodology to make use of spatial coherence

Tweaking the training phase, we manage to turn a very poor classification into a good one, close to that obtained by classifying in the same training region, however, several classifi-

Figure 5 – Methodology Overview



cation errors would clearly benefit from neighborhood information. A pixel is clearly less likely to be a sample of coffee plantation if there are none or few coffee pixels nearby. Information from neighbor pixels is also important because coffee patches form patterns in the images due to row spacing and the similar height of coffee plants. This pattern is clearly different from forest patches (Figure 6), which is most common source for false positives in the region of interest. Therefore, textural classification seems a promising tool.

Textural classification for coffee has been done for surface texture, i.e. canopy height variation, using SAR data (SILVA et al., 2009), and also done for image texture, i.e. variation of intensity in pixels (LELONG; THONG-CHANE, 2003; TSAI; CHEN, 2017). We think such methods still need further work, as there are many non trivial ways for representing texture in useful ways for classification. It is very dependent on image resolution, size of analyzed windows, sun/shadow arrangements and land slope (BAETA et al., 2017). Spatial coherence is also the principle behind object based classification methods such as done by Wang et al.

Figure 6 – Coffee plantation (upper region, delimited by a polygon) and forest (lower region) create distinct shadow patterns on Sentinel-2 10m resolution images.

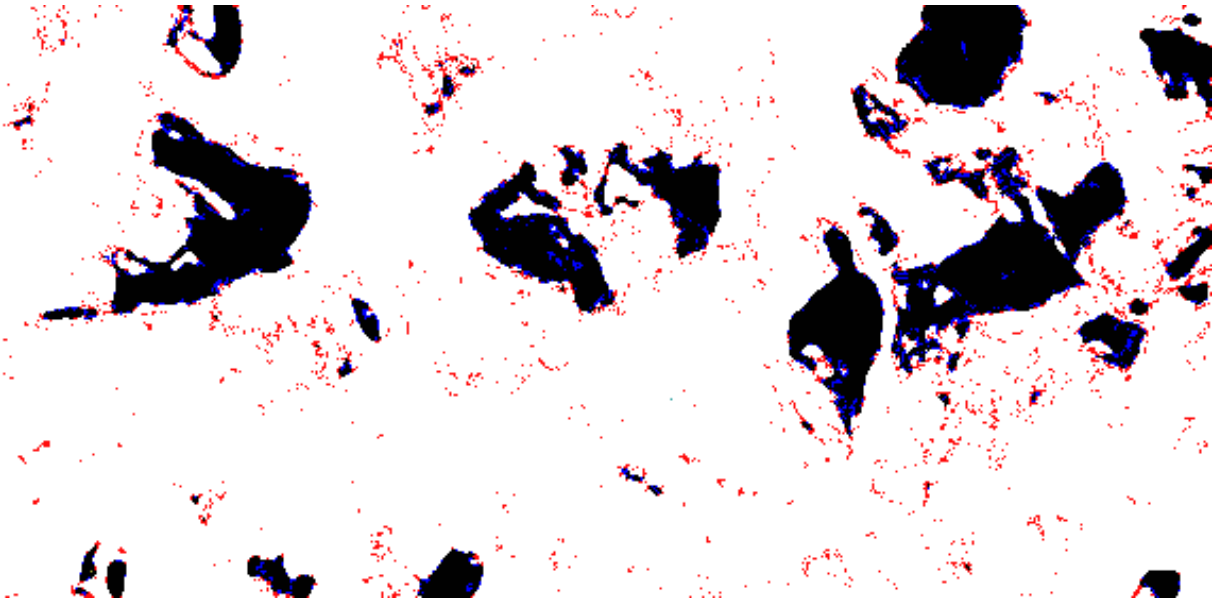
(WANG et al., 2018). We propose a simpler, direct way to put spatial information to use in the classification, without need of textural analysis.

The polygons for coffee plantations were rasterized, creating a raster image where there are ones are in place of the plantations and zeroes elsewhere. Because we had two classes for coffee, two images were created, one for each class. These images were then convoluted, so that each pixel gets the sum of its neighbors, creating information about how many nearby pixels are also from the same coffee class. We used a 5x5 pixel neighborhood.

The neighborhood information is clearly beneficial to the classifier, but to use it for classification one would need to know the classification output before the classification. To get around this problem, the classification was done in two steps. In the first step, classification is done without neighborhood information. The resulting classification is deemed an intermediary classification and used to create neighborhood information in the classification area. In the second step training is done again, with neighborhood information and classification is done again, but this time using neighborhood information from the intermediary classification. The classification done on the second step does not contains the salt-and-pepper effect of single,

misclassified pixels, borders are more accurate and holes inside plantations are smaller. The neighborhood information creates an eroding effect on false positives while have a dilating effect on false negatives, as seen in Figure 7.

Figure 7 – Effect of neighborhood for classification improvement - red pixels were removed from coffee classes and blue pixels were inserted.



Neighborhood information may be used in an iterative manner because subsequent classifications tend to be better than the previous. In our test region, accuracy increased from 92.4% to 94.3% with one iteration (adding neighborhood information), and then to 94.4% with two iterations (neighborhood updated). Such a marginal enhancement, however, suggests that using it as an iterative method may not be worth the extra computation. Applying neighborhood information creates more continuous areas, feasible for creation of polygons.

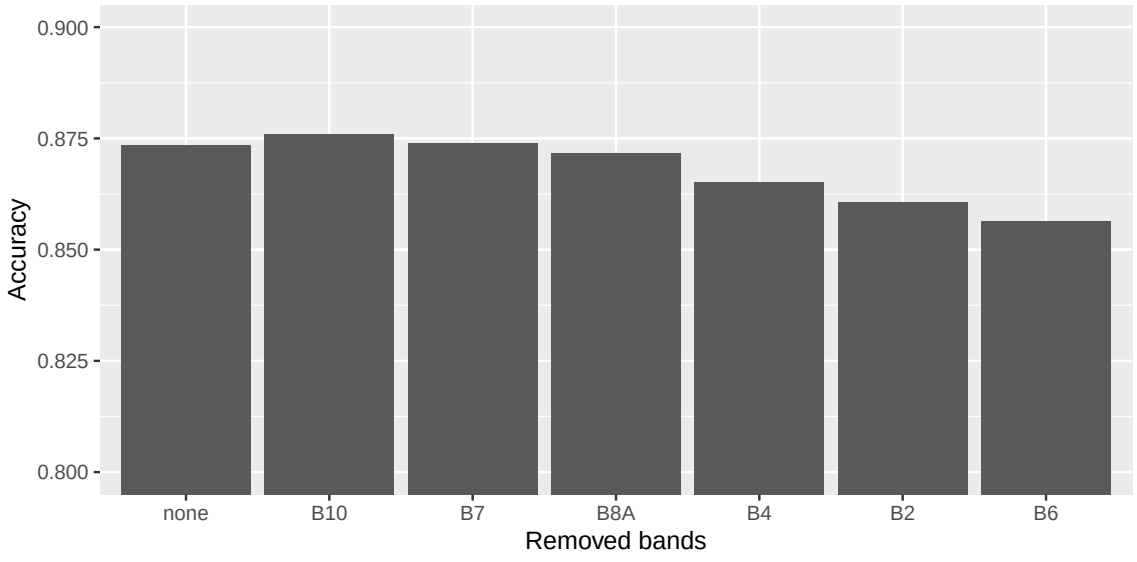
5 Results

Performing classification on the same area produced the highest accuracy. The loss of accuracy noticed when classifying on a different area could be mitigated using several techniques, including our novel spatial coherence modification to the training data. Random Forest classification has shown to be very robust, using large amounts of training samples, which simplify the sampling process by including as much different samples as possible. However, balancing the percentage of coffee and non-coffee samples shown to be important in training.

Figure 8 – Effect of second iteration on neighborhood data - red pixels were removed from coffee classes and blue pixels were inserted.

Removing bands form the training data is very dependent on the training data, removing a few of the less important bands, according to the internal Random Forest importance, we found that removing band 10 from training data improved accuracy by 0.3%. Changes in accuracy when removing the less important bands, according the importance classification from the previous set of bands, showed a decrease in accuracy after removing band 10 (Figure 9). In similar experimentation, with different scenarios than reported here, removing up to the 3rd less important bands, resulted in marginal classification improvement.

Figure 9 – Changes in accuracy when removing less important bands. Be aware that the y axis does not start at zero to enhance the differences. The x axis is cumulative, meaning that in each column, all bands removed previously are still absent from data.



Removing further bands decreased accuracy. Adding NDVI and NDWI bands accounted for another 0.3% increase in accuracy. Sentinel-2 Band 1 (Coastal aerosol 0.44μm) was surprisingly ranked first in importance for coffee mapping in every run.

Proportion of samples of different classes is important in the accuracy. Increasing the number of training samples from 16,605 (11.6% of all samples) to 35,510 (21.9%) resulted in 4.74% increase in overall accuracy and was the most useful improvement for classification.

Our novel neighborhood data increased accuracy in 1.9% for one iteration and 1.95% for two iterations, creating mapped regions more likely to be represented as coffee plantation polygons.

The loss of precision observed for classifying a region different from the training region could be recovered by using the techniques described. Table 2 is a summary of precision evolution as the classification was enhanced.

Table 2 – Timeline of classification precision.

Description	Precision	Difference
Training and classifying on Training Region 1, 13 bands	0.9694	–
Training on TR1, classifying on CR, 13 bands	0.8734	-0.0960
Training on TR1, classifying on CR, 12 bands	0.8760	+0.0026
Training on TR1 and TR2, classifying on CR, 12 bands	0.8931	+0.0171
Training on TR1, TR2 and TR3, classifying on CR	0.9234	+0.0303
Added NDVI and NDWI to data	0.9242	+0.0008
Added Neighborhood data	0.9429	+0.0187
Second iteration of neighborhood data	0.9437	+0.0008

6 Discussion

We noticed significant difference in accuracy when using the classification model for one region to another. Since both regions were near each other, on the same satellite image, for classifying similar crops, we expected similar results. This is a warning for giving too much importance on reported accuracy rates, on classification work. Not only classification varies a lot depending on how the training samples are chosen, but how also on also on how subsets are chosen.

Using Sentinel-2 data, the band 10 was ranked less important for coffee classification. Removing less important bands from training should be done iteratively, as the importance of

the remaining bands may change on the new model. We noticed that removing less important bands quickly exhausts its benefits. Removing one band was best for this work. We performed other classification tests with more coffee classes that benefit from removing at most three bands from the training data. Since improvement was small it seems that the general rule is that Random Forest classification benefits from having more spectral bands. We have no explanation for the consistent first place in importance rank for band 1 and believe this merits further investigation.

Previous work investigating how to sample data for training seem to be based on classes with as little noise as possible. We believe that is unpractical for large scale classification, which we should be aiming when implementing satellite data classification. The potential of generation of large mappings is the reason for investing in satellite classification and there is no avoiding large amounts of noise. Random Forest classification is clearly capable of dealing with such noise, but the samples used for training are still play a most significant role in accuracy. Classes with lots of noise must have a large number of samples, but with proportion in mind as to not harm other classes.

7 Conclusion

Coffee mapping using Sentinel-2 data and Random Forests classification algorithm achieves a good accuracy, comparable with what is been reported in the literature. Although accuracy may vary significantly depending on training data, classification may be fine tuned using simple methodologies. The disadvantages of pixel based classification may be mitigated using spatial coherence in the training.

References

BAETA, R. et al. Learning deep features on multiple scales for coffee crop recognition. In: **2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2017. p. 262–268.

BOURGOIN, C. et al. Assessing the ecological vulnerability of forest landscape to agricultural frontier expansion in the central highlands of vietnam. **International Journal of Applied Earth Observation and Geoinformation**, v. 84, p. 101958, 2020. ISSN 1569-8432. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0303243419307202>.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

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 International**, Taylor & Francis, v. 32, n. 7, p. 759–776, 2017.

ESCOBAR-LÓPEZ, A. et al. Identifying coffee agroforestry system types using multitemporal sentinel-2 data and auxiliary information. **Remote Sensing**, v. 14, n. 16, 2022. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/14/16/3847>.

FINER, M. et al. Combating deforestation: From satellite to intervention. **Science**, v. 360, n. 6395, p. 1303–1305, 2018. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aat1203>.

KELLEY, L. C.; PITCHER, L.; BACON, C. Using google earth engine to map complex shade-grown coffee landscapes in northern nicaragua. **Remote Sensing**, v. 10, n. 6, 2018. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/10/6/952>.

LELONG, C. C. D.; THONG-CHANE, A. Application of textural analysis on very high resolution panchromatic images to map coffee orchards in uganda. In: **2003 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2003, Toulouse, France, July 21-15, 2003**. IEEE, 2003. p. 1007–1009. Disponível em: <https://doi.org/10.1109/IGARSS.2003.1293994>.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. **Remote Sensing**, 2019.

SILVA, W. F. et al. Discrimination of agricultural crops in a tropical semi-arid region of brazil based on l-band polarimetric airborne sar data. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 5, p. 458–463, 2009. ISSN 0924-2716. Theme Issue: Mapping with SAR: Techniques and Applications. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0924271608000932>.

TSAI, D.-M.; CHEN, W.-L. Coffee plantation area recognition in satellite images using fourier transform. **Computers and Electronics in Agriculture**, v. 135, p. 115–127, 2017. ISSN 0168-1699.

WANG, W. et al. Improving object-based land use/cover classification from medium resolution imagery by markov chain geostatistical post-classification. **Land**, v. 7, n. 1, 2018. ISSN 2073-445X. Disponível em: <https://www.mdpi.com/2073-445X/7/1/31>.

APPENDIX A – Source code for classification

The source code presented here, in R, using `terra` and `ranger` libraries, creates a mapping from Sentinel-2 data, manual classification polygons and area delimiters. Data files are not included.

```
#!/usr/bin/env Rscript

# This script creates a classification model using 3 training regions, without band 10, adding NDVI and NDWI, and uses two
# passes of neighborhood information.

# Author: Bruno de Oliveira Schneider - UFLA, 2023

cat('\n=====================================================================\n')
cat('3 TRs, B10 drop, +NDVI +NDWI, 2 neighborhood passes.\n')

library(terra)

# Load training regions delimiters
trainingRegion1 <- vect('regiao-treinamento.shp')
trainingRegion2 <- vect('regiao-treinamento2.shp')
trainingRegion3 <- vect('regiao-treinamento3.shp')
# Load classification region delimiter
classifRegion <- vect('regiao-classificacao.shp')

# Load Sentinel-2 data (higher/reference resolution first)
b2  <- rast('T23KMS_20220625T131301_B02.jp2')
b2T <- crop(b2, trainingRegion1)
# Other data with the same higher resolution
b3  <- rast('T23KMS_20220625T131301_B03.jp2')
b4  <- rast('T23KMS_20220625T131301_B04.jp2')
b8  <- rast('T23KMS_20220625T131301_B08.jp2')

# Load bands that need to be resampled
b1  <- rast('T23KMS_20220625T131301_B01.jp2')
b5  <- rast('T23KMS_20220625T131301_B05.jp2')
b6  <- rast('T23KMS_20220625T131301_B06.jp2')
b7  <- rast('T23KMS_20220625T131301_B07.jp2')
b8a <- rast('T23KMS_20220625T131301_B8A.jp2')
b9  <- rast('T23KMS_20220625T131301_B09.jp2')
b10 <- rast('T23KMS_20220625T131301_B10.jp2')
b11 <- rast('T23KMS_20220625T131301_B11.jp2')
b12 <- rast('T23KMS_20220625T131301_B12.jp2')

# Create multi spectral image
bNIR <- crop(b8, trainingRegion1)
bRed <- crop(b4, trainingRegion1)
bMIR <- resample(b12, b2T)
bNDVI <- (bNIR-bRed) / (bNIR+bRed)
names(bNDVI) <- c('NDVI')
bNDWI <- (bNIR-bMIR) / (bNIR+bMIR)
names(bNDWI) <- c('NDWI')
multiTraining1 <- c(resample(b1, b2T),
                    b2T,
                    crop(b3, trainingRegion1),
                    bRed,
                    resample(b5, b2T),
                    resample(b6, b2T),
                    resample(b7, b2T),
                    bNIR,
                    resample(b8a, b2T),
                    resample(b9, b2T),
                    #resample(b10, b2T),
                    resample(b11, b2T),
                    bMIR)
```

```
names(multiTraining1) <- c('coastal', 'blue', 'green', 'red', 'RedE5', 'RedE6', 'RedE7', 'NIR', 'NNIR', 'wVapor',
                           'SWIR_11', 'SWIR_12')

# Create vegetation indices NDVI and NDWI
b2t2 <- crop(b2, trainingRegion2)
bNIR2 <- crop(b8, trainingRegion2)
bRed2 <- crop(b4, trainingRegion2)
bMIR2 <- resample(b12, b2t2)
bNDVI2 <- (bNIR2-bRed2) / (bNIR2+bRed2)
names(bNDVI2) <- c('NDVI')
bNDWI2 <- (bNIR2-bMIR2) / (bNIR2+bMIR2)
names(bNDWI2) <- c('NDWI')
multiTraining2 <- c(resample(b1, b2t2),
                b2t2,
                crop(b3, trainingRegion2),
                bRed2,
                resample(b5, b2t2),
                resample(b6, b2t2),
                resample(b7, b2t2),
                bNIR2,
                resample(b8a, b2t2),
                resample(b9, b2t2),
                #resample(b10, b2T),
                resample(b11, b2t2),
                bMIR2)
names(multiTraining2) <- c('coastal', 'blue', 'green', 'red', 'RedE5', 'RedE6', 'RedE7', 'NIR', 'NNIR', 'wVapor',
                           'SWIR_11', 'SWIR_12')
b2t3 <- crop(b2, trainingRegion3)
bNIR3 <- crop(b8, trainingRegion3)
bRed3 <- crop(b4, trainingRegion3)
bMIR3 <- resample(b12, b2t3)
bNDVI3 <- (bNIR3-bRed3) / (bNIR3+bRed3)
names(bNDVI3) <- c('NDVI')
bNDWI3 <- (bNIR3-bMIR3) / (bNIR3+bMIR3)
names(bNDWI3) <- c('NDWI')
multiTraining3 <- c(resample(b1, b2t3),
                b2t3,
                crop(b3, trainingRegion3),
                bRed3,
                resample(b5, b2t3),
                resample(b6, b2t3),
                resample(b7, b2t3),
                bNIR3,
                resample(b8a, b2t3),
                resample(b9, b2t3),
                #resample(b10, b2t3),
                resample(b11, b2t3),
                bMIR3)
names(multiTraining3) <- c('coastal', 'blue', 'green', 'red', 'RedE5', 'RedE6', 'RedE7', 'NIR', 'NNIR', 'wVapor',
                           'SWIR_11', 'SWIR_12')


# Classification region data
b2C <- crop(b2, classifRegion)
bNIRC <- crop(b8, classifRegion)
bRedC <- crop(b4, classifRegion)
bMIRC <- resample(b12, b2C)
bNDVIC <- (bNIRC-bRedC) / (bNIRC+bRedC)
names(bNDVIC) <- c('NDVI')
bNDWIC <- (bNIRC-bMIRC) / (bNIRC+bMIRC)
names(bNDWIC) <- c('NDWI')
multiClassif <- c(resample(b1, b2C),
                b2C,
                crop(b3, classifRegion),
                bRedC,
                resample(b5, b2C),
```

```
                resample(b6, b2C),
                resample(b7, b2C),
                bNIRC,
                resample(b8a, b2C),
                resample(b9, b2C),
                #resample(b10, b2C),
                resample(b11, b2C),
                bMIRC)
names(multiClassif) <- c('coastal', 'blue', 'green', 'red', 'RedE5', 'RedE6', 'RedE7', 'NIR', 'NNIR', 'wVapor',
                         'SWIR_11', 'SWIR_12')

# Plot RGB image from classification image if intended
#plotRGB(multiTraining1, 4,3,2)
#img <- multiClassif[[c(4,3,2)]] # copy from the original set

# Load verified EMPAMIG/EMATER coffee polygons
cropPolys <- vect('cafe-lavras-edt.shp')
trainingPolys1 <- crop(cropPolys, trainingRegion1)
trainingPolys1$AREA_ha <- NULL # remove area column
trainingPolys2 <- crop(cropPolys, trainingRegion2)
trainingPolys2$AREA_ha <- NULL # remove area column
trainingPolys3 <- crop(cropPolys, trainingRegion3)
trainingPolys3$AREA_ha <- NULL # remove area column
classifPolys <- crop(cropPolys, classifRegion)
classifPolys$AREA_ha <- NULL # remove area column

# Rasterize crop polygons / balancing coffee samples
rasterClasses <- rasterize(trainingPolys1, b2T, field='Categoria', background=0)
names(rasterClasses) <- c('class')
trainingData1 <- as.data.frame(c(rasterClasses, multiTraining1, bNDVI, bNDWI))
rasterClasses2 <- rasterize(trainingPolys2, b2t2, field='Categoria', background=0)
names(rasterClasses2) <- c('class')
trainingData2 <- as.data.frame(c(rasterClasses2, multiTraining2, bNDVI2, bNDWI2))
trainingData2 <- trainingData2[trainingData2$class > 0, ] # filter non-coffee samples
rasterClasses3 <- rasterize(trainingPolys3, b2t3, field='Categoria', background=0)
names(rasterClasses3) <- c('class')
trainingData3 <- as.data.frame(c(rasterClasses3, multiTraining3, bNDVI3, bNDWI3))
trainingData3 <- trainingData3[trainingData3$class > 0, ] # filter non-coffee samples
fullTrainingData <- rbind(trainingData1, rbind(trainingData2, trainingData3))
cat(paste(nrow(fullTrainingData), 'samples used in training.\n'))
cat('Number of samples for each class:')
print(table(fullTrainingData$class))

# Create dataframe for classification
classifData <- as.data.frame(c(multiClassif, bNDVIC, bNDWIC))

# Define inputs and output
predictors <- c(names(multiTraining1), 'NDVI', 'NDWI')
form <- as.formula(paste('class ~', paste(predictors, collapse='+'))) #warning: names must not contain spaces

# Create classification model (this may take some time...)
library("ranger")
classifModel <- ranger(form, data=fullTrainingData, importance='impurity', classification=TRUE)
cat('Band importance ranking:\n')
print(classifModel$variable.importance[order(classifModel$variable.importance)])
cat(paste('Classification error (OOB):', classifModel$prediction.error, '\n'))

# Classify!
pred <- predict(classifModel, data=classifData, type='response')

# Accuracy function
ovAcc <- function(conmat) {
    # number of total cases/samples
    n = sum(conmat)
    # number of correctly classified cases per class
    diag = diag(conmat)
```

```
    # Overall Accuracy
    OA = sum(diag) / n
    # observed (true) cases per class
    rowsums = apply(conmat, 1, sum)
    p = rowsums / n
    # predicted cases per class
    colsums = apply(conmat, 2, sum)
    q = colsums / n
    expAccuracy = sum(p*q)
    kappa = (OA - expAccuracy) / (1 - expAccuracy)
    # Producer accuracy
    PA <- diag / colsums
    # User accuracy
    UA <- diag / rowsums
    outAcc <- data.frame(producerAccuracy = PA, userAccuracy = UA)
    #print(outAcc)

    global_acc = data.frame(overallAccuracy=OA, overallKappa=kappa)
    #print(global_acc)
    cat('Producer accuracy for each class:\n')
    print(PA)
    cat(paste('Overall accuracy:', format(OA, digits=4), 'Kappa:', format(kappa, digits=4), '\n'))
    # Based from: http://gsp.humboldt.edu/olm/Courses/GSP_216/lessons/accuracy/metrics.html
}


# Create image from classification
rastClassif1 <- rast(ncols=ncol(b2C), nrows=nrow(b2C), nlyrs=1, crs=crs(b2C), extent=ext(b2C))
values(rastClassif1) <- pred$predictions
names(rastClassif1) <- c('class')

# Save to file, if intended. Supposing 2 classes, generates values up to 200.
#writeRaster(rastClassif1*100, 'classificacao.png', overwrite=TRUE)
#plot(rastClassif1)
#plot(classifPolys, add=TRUE)

# Rasterize coffee polygons to generation a true classification
trueClassif <- terra::rasterize(classifPolys, b2C, field='Categoria', background=0)

# Confusion Matrix
realFactors <- factor(values(trueClassif), levels=c(0:2))
compFactors <- factor(pred$predictions, levels=c(0:2))
conMat <- table(compFactors, realFactors)
ovAcc(conMat)

# Generate raster from first neighborhood information. TR1.
cat('Computing neighborhood data\n')
neighborhood1 <- terra::rasterize(trainingPolys1[trainingPolys1$Categoria==1], b2T, field=1, background=0)
neighborhood1 <- terra::focal(neighborhood1, w=5, fun='sum', fillvalue=0)
names(neighborhood1) <- c('nbh1')
neighborhood2 <- terra::rasterize(trainingPolys1[trainingPolys1$Categoria==2], b2T, field=1, background=0)
neighborhood2 <- terra::focal(neighborhood2, w=5, fun='sum', fillvalue=0)
names(neighborhood2) <- c('nbh2')
multiTraining1 <- c(multiTraining1, neighborhood1, neighborhood2)
#plot(neighborhood1)
#print("Neighborhood histogram:")
#print(table(as.vector(neighborhood1)))

# Generate raster from first neighborhood information. TR2.
neighborhood1 <- terra::rasterize(trainingPolys2[trainingPolys2$Categoria==1], b2t2, field=1, background=0)
neighborhood1 <- terra::focal(neighborhood1, w=5, fun='sum', fillvalue=0)
names(neighborhood1) <- c('nbh1')
neighborhood2 <- terra::rasterize(trainingPolys2[trainingPolys2$Categoria==2], b2t2, field=1, background=0)
neighborhood2 <- terra::focal(neighborhood2, w=5, fun='sum', fillvalue=0)
names(neighborhood2) <- c('nbh2')
multiTraining2 <- c(multiTraining2, neighborhood1, neighborhood2)
```

```
#plot(neighborhood1)

# Generate raster from first neighborhood information. TR3.
neighborhood1 <- terra::rasterize(trainingPolys3[trainingPolys3$Categoria==1], b2t3, field=1, background=0)
neighborhood1 <- terra::focal(neighborhood1, w=5, fun='sum', fillvalue=0)
names(neighborhood1) <- c('nbh1')
neighborhood2 <- terra::rasterize(trainingPolys3[trainingPolys3$Categoria==2], b2t3, field=1, background=0)
neighborhood2 <- terra::focal(neighborhood2, w=5, fun='sum', fillvalue=0)
names(neighborhood2) <- c('nbh2')
multiTraining3 <- c(multiTraining3, neighborhood1, neighborhood2)
#plot(neighborhood1)

# Create new classification model using neighborhood data (this may take some time...)
trainingData1 <- as.data.frame(c(rasterClasses, multiTraining1, bNDVI, bNDWI))
trainingData2 <- as.data.frame(c(rasterClasses2, multiTraining2, bNDVI2, bNDWI2))
trainingData2 <- trainingData2[trainingData2$class>0, ] # filtrar a classe 0
trainingData3 <- as.data.frame(c(rasterClasses3, multiTraining3, bNDVI3, bNDWI3))
trainingData3 <- trainingData3[trainingData3$class>0, ] # filtrar a classe 0
# Join all data
fullTrainingData <- rbind(trainingData1, rbind(trainingData2, trainingData3))
predictors <- c(names(multiTraining1), 'NDVI', 'NDWI')
form <- as.formula(paste('class ~', paste(predictors, collapse='+')))
cat('Creating new classification model\n')
classifModel <- ranger(form, data=fullTrainingData, importance="impurity", classification=TRUE)
cat(paste('Classification error (OOB):', classifModel$prediction.error, '\n'))

# Create neighborhood data for CR
neighborhood1 <- rastClassif1
neighborhood1[neighborhood1$class==2] = 0 # filter class 2
neighborhood1 <- terra::focal(neighborhood1, w=5, fun='sum', fillvalue=0)
names(neighborhood1) <- c('nbh1')
neighborhood2 <- rastClassif1
neighborhood2[neighborhood2$class==1] = 0 # filter class 1
neighborhood2[neighborhood2$class==2] = 1
neighborhood2 <- terra::focal(neighborhood2, w=5, fun='sum', fillvalue=0)
names(neighborhood2) <- c('nbh2')

classifData <- as.data.frame(c(multiClassif, neighborhood1, neighborhood2, bNDVIC, bNDWIC))
cat('Second mapping:\n')
pred <- predict(classifModel, data=classifData, type='response')
rastClassif2 <- rastClassif1
values(rastClassif2) <- pred$prediction
#writeRaster(rastClassif2*51, 'classificacao.png', overwrite=TRUE)

# Update confusion matrix
compFactors <- factor(pred$predictions, levels=c(0:2))
conMat <- table(compFactors, realFactors)
ovAcc(conMat)
#writeRaster(rastClassif2*100, 'classificacao-it1.png', overwrite=TRUE)

cat('\nSecond neighborhood iteration\n')
neighborhood1 <- rastClassif2
neighborhood1[neighborhood1$class==2] = 0 # filter class 2
neighborhood1 <- terra::focal(neighborhood1, w=5, fun='sum', fillvalue=0)
names(neighborhood1) <- c('nbh1')
neighborhood2 <- rastClassif2
neighborhood2[neighborhood2$class==1] = 0 # filter class 1
neighborhood2[neighborhood2$class==2] = 1
neighborhood2 <- terra::focal(neighborhood2, w=5, fun='sum', fillvalue=0)
names(neighborhood2) <- c('nbh2')

classifData <- as.data.frame(c(multiClassif, neighborhood1, neighborhood2, bNDVIC, bNDWIC))
cat('Third mapping:\n')
pred <- predict(classifModel, data=classifData, type='response')
rastClassif3 <- rastClassif1
values(rastClassif3) <- pred$prediction
```

```
compFactors <- factor(pred$predictions, levels=c(0:2))
conMat <- table(compFactors, realFactors)
ovAcc(conMat)
#writeRaster(rastClassif3*100, 'classificacao-it2.png', overwrite=TRUE)
```