

ALINE BARROCA MARRA

**DETECÇÃO E MAPEAMENTO DE NEMATOIDES NA CULTURA
CAFEEIRA POR MEIO DE IMAGENS MULTIESPECTRAIS DO
MSI/SENTINEL-2 E CLASSIFICAÇÃO BASEADA EM APRENDIZADO
DE MÁQUINA**



PRESIDENTE PRUDENTE
2020



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Faculdade de Ciências e Tecnologia
Campus de Presidente Prudente
Programa de Pós-Graduação em Ciências Cartográficas

ALINE BARROCA MARRA

**DETECÇÃO E MAPEAMENTO DE NEMATOIDES NA CULTURA
CAFEIEIRA POR MEIO DE IMAGENS MULTIESPECTRAIS DO
MSI/SENTINEL-2 E CLASSIFICAÇÃO BASEADA EM APRENDIZADO
DE MÁQUINA**

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Cartográficas da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista “Júlio de Mesquita Filho”.

Orientadora:
Dra. Maria de Lourdes Bueno Trindade Galo

Coorientador:
Dr. George Deroco Martins

PRESIDENTE PRUDENTE
2020

FICHA CATALOGRÁFICA

M358d

Marra, Aline Barroca

Detecção e mapeamento de nematoides na cultura cafeeira por meio de imagens multiespectrais do MSI/Sentinel-2 e classificação baseada em aprendizado de máquina /

Aline Barroca Barra. -- Presidente Prudente, 2020

112 f. : il., tabs., mapas

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientadora: Maria de Lourdes Bueno Trindade Galo

Coorientador: George Deroco Martins

1. Cultura cafeeira. 2. Detecção de doenças. 3. Imagens multiespectrais. 4. Aprendizado de máquina. 5. Mapeamento. I. Título.

CERTIFICADO DE APROVAÇÃO

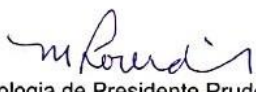
TÍTULO DA DISSERTAÇÃO: Detecção e mapeamento de nematóides na cultura cafeeira por meio de imagens multiespectrais do MSI/SENTINEL-2 e classificação baseada em aprendizado de máquina

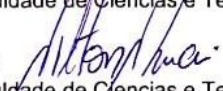
AUTORA: ALINE BARROCA MARRA

ORIENTADORA: MARIA DE LOURDES BUENO TRINDADE GALO

COORIENTADOR: GEORGE DEROCO MARTINS

Aprovada como parte das exigências para obtenção do Título de Mestra em CIÊNCIAS CARTOGRÁFICAS, área: Aquisição, Análise e Representação de Informações Espaciais pela Comissão Examinadora:

Profa. Dra. MARIA DE LOURDES BUENO TRINDADE GALO 
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente

Prof. Dr. NILTON NOBUHIRO IMAI 
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente

Prof. Dr. HELIO RICARDO SILVA 
Departamento de Fitossanidade, Engenharia Rural e Solos / Faculdade de Engenharia de Ilha Solteira

VIDEOCONFERÊNCIA

Presidente Prudente, 04 de março de 2020

*Aos meus pais, Iderval e Neuza e
irmãs Daniele, Luana e Soellyn.*

AGRADECIMENTOS

Aos meus pais Iderval e Neuza por sempre me apoiarem em todas as minhas decisões e pelo incentivo nos momentos difíceis.

Às minhas irmãs dani, lu e su por estarem sempre presente me incentivando e me acompanhando em cada passo. Em especial à dani, que além de minha irmã, foi minha professora, sempre achando um tempinho para me acompanhar passo a passo e me ajudar em tudo que eu precisava.

Aos meus sobrinhos Breno, Alisson e Júlia por me proporcionarem alegria e muito amor.

À minha orientadora Prof. Dra. Maria de Lourdes pela dedicação, confiança e pelas aulas de sabedoria e conhecimento nas reuniões. Sempre muito paciente e atenciosa nas minhas dúvidas.

Ao meu co-orientador Prof. Dr. George Deroco Martins da Universidade Federal de Uberlândia (UFU), pela atenção e contribuição essencial para realização desta dissertação.

Aos membros da banca examinadora: Prof. Dra. Maria de Lourdes Bueno Trindade Galo (UNESP), Prof. Dr. Nilton Nobuhiro Imai (UNESP) e Prof. Dr. Hélio Ricardo Silva (UNESP), pela participação e valiosas discussões e sugestões.

Aos colegas do programa Franciele, Viviane e Luiz Eduardo pelo apoio e ajuda.

Ao Programa de Pós-Graduação em Ciências Cartográficas (PPGCC) e à Universidade Estadual Paulista (UNESP) pela infraestrutura fornecida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

O Brasil é responsável por grande parte da produção mundial de café. Assim como em outras culturas agrícolas, a infecção por parasitas é comum em cafeeiros. Nematoides são parasitas do solo e do sistema radicular da planta que afetam a absorção de água e de nutrientes, podendo causar estresse hídrico e comprometer o desenvolvimento da parte aérea. As plantas infectadas por nematoides tem seu comportamento espectral alterado em relação às plantas saudáveis. Especificamente quanto à região espectral do infravermelho de ondas curtas (SWIR - *short wave infrared*), região sensível a variações no conteúdo de água da planta, considera-se que a ocorrência de nematoides no cafeeiro possa ser detectada com maior eficiência. Nesse sentido, o objetivo desta pesquisa foi estudar e avaliar o potencial das imagens multiespectrais do sistema orbital Sentinel-2 na detecção de áreas parasitadas por nematoides em plantação cafeeira, bem como desenvolver um método para mapeamento de áreas infectadas pelo parasita, baseado em uma abordagem de aprendizado de máquina. O estudo foi desenvolvido em uma área cafeeira do Cerrado Mineiro na qual foi constatada e registrada a ocorrência de talhões infectados por nematoides. A fim de avaliar o potencial das bandas espectrais originais e de diferentes tipos de índices de vegetação na detecção do parasita, áreas saudáveis e infectadas por nematoides foram identificadas e utilizadas como referência para a realização de dois experimentos. Esses experimentos auxiliaram no processo de seleção dos atributos mais significativos para uma comparação do desempenho dos classificadores Florestas Aleatórias (FA) e Redes Neurais Artificiais (RNAs). Os resultados indicaram que o FA apresentou melhores indicadores de acurácia e menor tempo de processamento. Quanto aos mapas gerados após a classificação, os dois classificadores apresentaram produtos semelhantes, porém, no que se refere à definição das áreas onde a atribuição de um pixel a uma classe foi mais incerta, a RNA teve desempenho superior ao FA. Assim, o estudo desenvolvido mostrou que a utilização de bandas na região espectral do (SWIR) do Sentinel-2, ofereceu eficiência na detecção de nematoides em plantação de café, por ser uma região mais sensível ao teor de umidade da planta, permitindo uma melhor investigação da doença devido aos efeitos causados pelo parasita. Destaca-se também a contribuição de uma das bandas do vermelho limítrofe, a Red Edge 3 do sistema Sentinel-2, no aumento do poder dessa detecção, em conjunto com a região espectral do SWIR. O estudo também mostrou que o mapeamento pode trazer indícios de localizações com a ocorrência de nematoides no cafeeiro, se tornando uma ferramenta de ajuda para o fornecimento informações fundamentais no aumento da eficiência do manejo e para alcançar as expectativas de produção agrícola.

Palavras chave: detecção de doenças, cultura de café, Sentinel-2, nematoides, classificação baseada em aprendizado de máquina, mapeamento.

ABSTRACT

Brazil is responsible for much of the world's coffee production. As with other crops, parasite infection is common in coffee crops. Nematodes are soil and root system parasites that affect plant water and nutrient absorption, which can cause water stress and compromising the aerial part development. Plants infected by nematodes have their spectral behavior altered in relation to healthy plants. Specifically regarding to the spectral region of the short-wave infrared (SWIR), a region sensitive to variations in the water content of the plant, it is considered that the occurrence of nematodes in the coffee crop can be detected with greater efficiency. In this sense, the objective of this research was to study and evaluate the potential of multispectral images of the MSI/Sentinel-2 orbital system in the detection of nematode-infected coffee plantation, as well as to develop a method for mapping areas infected by the parasite, based on a machine learning approach. The study was carried out in a coffee area of Cerrado Mineiro where it was found and recorded the occurrence of nematode-infected fields. In order to evaluate the potential of original spectral bands and different types of vegetation indices for the parasite detection, healthy and nematode infected areas were identified and used as references for the realization of two experiments. These experiments helped in the selection process of the most significant features for a comparison of the performance of the classifiers Random Forests (RF) and Artificial Neural Networks (ANNs). The results indicated that the RF showed better indicators of accuracy and shorter processing time. As for the maps generated after the classification, the two classifiers presented similar products, however, with regard to the definition of areas where the assignment of a pixel to a class was more uncertain, the ANN performed better than the RF. Thus, the study developed showed that the use of the SWIR spectral region of the Sentinel-2 system, offered efficiency in the detection of nematodes in coffee plantations, as it is a region more sensitive to the moisture content of the plant, allowing a better investigation of the disease due to the effects caused by the parasite. Also noteworthy is the contribution of one of the Red Edge bands, the Red Edge 3 of the Sentinel-2 system, in increasing the power of this detection, together with the spectral region of the SWIR. The study also showed that the mapping can bring indications of locations with the occurrence of nematodes in the coffee tree, becoming a tool to help provide fundamental information to increase the efficiency of management and to reach the expectations of agricultural production.

Keywords: disease detection, coffee plantation, Sentinel-2, nematodes, machine learning classification, mapping.

SUMÁRIO

1. INTRODUÇÃO	10
2. OBJETIVOS.....	13
2.1 Objetivo geral	13
2.2 Objetivos específicos.....	13
3. JUSTIFICATIVA.....	14
4. FUNDAMENTAÇÃO TEÓRICA	16
4.1 Fenologia do cafeeiro	16
4.2 Nematoides na plantação de café.....	17
4.3 Sensoriamento remoto da vegetação	21
4.3.1 Interação da radiação eletromagnética com a vegetação	21
4.3.2 Comportamento Espectral da Vegetação	23
4.4 Índices espectrais	26
4.5 Imagem de textura	29
4.5.1 Construção da matriz de co-ocorrência	29
4.6 Classificação baseada em aprendizado de máquina.....	31
5. MATERIAIS E MÉTODOS.....	39
5.1 Materiais	39
5.1.1 Área de estudo	39
5.1.2 Dados de referência	41
5.1.3 Imagens multiespectrais MSI/Sentinel-2	43
5.1.4 Ferramentas Computacionais	44
5.2 Metodologia	46
5.2.1 Aquisição dos dados multiespectrais MSI/Sentinel-2	48
5.2.2 Pré-processamento das imagens	50
5.2.3 Geração dos índices de vegetação e imagens de textura	52
5.2.4 Seleção de atributos.....	53
5.2.4.1 Primeiro experimento	53
5.2.4.2 Segundo experimento	59
5.2.5 Classificação	60
5.2.5.1 Geração dos mapas resultantes da classificação	63
5.2.5.2 Validação dos dados classificados.....	63
6. RESULTADOS.....	66
6.1 Análise de atributos	66
6.1.1 Primeiro experimento	66

6.1.2 Segundo experimento	74
6.1.3 Definição dos conjuntos de atributos	80
6.2 Resultado da aplicação dos algoritmos de classificação	81
6.2.1 Desempenho dos algoritmos utilizados na classificação das imagens multitemporais	84
6.2.2 Mapas temáticos resultantes das classificações multitemporais	87
7. CONCLUSÕES E RECOMENDAÇÕES	100
REFERÊNCIAS	103

LISTA DE FIGURAS

Figura 1 - Esquematização das seis fases fenológicas do cafeeiro arábica, durante 24 meses ...	16
Figura 2 - Lavoura nova de café arábica implantada em área infestada por <i>M. paranaensis</i>	19
Figura 3 - Estrutura interna da planta.....	21
Figura 4 - Curva de reflectância típica de uma folha verde	23
Figura 5 - Comportamento espectral de folha verde e senescente	25
Figura 6 - Influência no conteúdo de água em folhas de milho	25
Figura 7 – Construção de uma matriz de co-ocorrência.....	30
Figura 8 - Estrutura de um neurônio artificial.....	33
Figura 9 - Modelo de rede neural com uma camada escondida	33
Figura 10 - Princípio do classificador florestas aleatórias	37
Figura 11 - Localização da área de estudo referente ao município de Monte Carmelo – MG....	39
Figura 12 – Chuva mensal média do município de Monte Carmelo – MG, referente aos anos de 2018 e 2019.....	40
Figura 13 – Temperatura mensal média do município de Monte Carmelo – MG, referente aos anos de 2018 e 2019.....	41
Figura 14 – Amostras de café infectado disponibilizadas pela UFU referente ao primeiro levantamento de campo.....	42
Figura 15 – Localização das amostras de café infectado e sadio do segundo conjunto de dados disponibilizados pela UFU	42
Figura 16 – Exemplo de talhões de café sadio (a) e café infectado (b).....	43
Figura 17 - Fluxograma geral da metodologia	48
Figura 18 – Inserção da imagem Sentinel-2A da área de estudos, sobreposta à visualização do <i>Google Earth</i>	49
Figura 19 - Processo de correção atmosférica do algoritmo <i>Sen2Cor</i>	51
Figura 20 – Recorte da imagem abrangendo os talhões de café em estudo (delimitados em vermelho)	52
Figura 21 - Exemplo de parte dos dados de treinamento em arquivo ASCII.....	54
Figura 22 - Localização da área teste, destacada a partir da área de estudos	56
Figura 23 - Exemplo do arquivo ARFF contendo parte dos dados de treinamento	56
Figura 24 – Exemplo de um arquivo formato ARFF a ser classificado	62
Figura 25 – Exemplo do arquivo de saída da classificação do WEKA.....	63
Figura 26 – Médias dos valores de reflectância de superfície dos pixels amostrados de café infectado e sadio obtidas das bandas espectrais do Sentinel-2 para a cena de julho de 2018	67
Figura 27 – Variação espectral nas médias dos valores de reflectância de superfície e seus respectivos desvios-padrão, obtidos para as amostras de café infectado (a) e sadio (b) nas bandas espectrais da cena de julho de 2018 do Sentinel-2.....	67

Figura 28 – Largura das bandas espectrais do Sentinel-2	69
Figura 29 – Médias dos índices de vegetação normalizados para amostras de café infectado e sadio, obtidas a partir da cena de julho de 2018 do Sentinel-2	70
Figura 30 – Médias dos índices de vegetação não-normalizados para amostras de café infectado e sadio, obtidas a partir da cena de julho de 2018 do Sentinel-2.....	70
Figura 31 – Matriz de correlação dos atributos.....	74
Figura 32 – Matriz de correlação dos atributos referente ao mês de janeiro/2019.....	75
Figura 33 – Matriz de correlação dos atributos referente ao mês de abril/2019	77
Figura 34 - Matriz de correlação dos atributos referente ao mês de julho/2018	79
Figura 35 - Matrizes de confusão geradas a partir dos dados de validação do classificador FA das cenas de (a) janeiro/2019, (b) abril/2019 e (c) julho/2018, em que CI = café infectado, CS = café sadio e SE = solo exposto	85
Figura 36 - Matrizes de confusão geradas a partir dos dados de validação do classificador RNA Artificiais das cenas de (a) janeiro/2019, (b) abril/2019 e (c) julho/2018, em que CI = café infectado, CS = café sadio e SE = solo exposto.	86
Figura 37 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de janeiro/2019	88
Figura 38 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de abril/2019.....	89
Figura 39 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de julho/2018	90
Figura 40 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a janeiro/2019.....	91
Figura 41 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a abril/2019	91
Figura 42 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a julho/2018.....	92
Figura 43 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem janeiro/2019	93
Figura 44 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem abril/2019	94
Figura 45 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem julho/2018	95
Figura 46 - Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a janeiro/2019.....	96
Figura 47 – Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a abril/2019	96
Figura 48 – Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a julho/2018.....	97

LISTA DE TABELAS

Tabela 1 – Formulação dos índices de vegetação utilizados. NIR (<i>Near Infrared</i>) se refere ao infravermelho próximo (comprimento de onda central 832,8 nm), RED ao vermelho (664,6 nm), SWIR1 (1613,7 nm) e SWIR2 (2202,4 nm) aos comprimentos de onda do infravermelho de ondas curtas 1 e 2 (<i>Short Wave Infrared</i>), RE1 ao <i>Red-Edge</i> 1 (704,1 nm) e GREEN ao canal verde (559,8 nm)	27
Tabela 2 - Variedades de café produzidos no município de Monte Carmelo - MG.....	40
Tabela 3 - Bandas espectrais do Sentinel-2.....	44
Tabela 4 – Número de pixels de treinamento amostrados por classe.....	60
Tabela 5: Médias e desvios-padrão da reflectância de superfície obtidos a partir das amostras de café infectado e sadio e a diferença absoluta entre as médias, para cada banda espectral do Sentinel-2	68
Tabela 6: Médias e desvios-padrão dos valores de índices de vegetação obtidos a partir das amostras de café infectado e sadio e a diferença absoluta entre as médias	71
Tabela 7: Valores M obtidos para os índices de vegetação normalizados e não-normalizados..	72
Tabela 8: <i>Ranking</i> de Importância dos atributos definido pelas métricas <i>Classifier Attribute Evaluation (ClassAE)</i> , <i>Gain Ratio Attribute Evaluation (GRAE)</i> e <i>Correlation Attribute Evaluation (CorrAE)</i>	73
Tabela 9 - <i>Ranking</i> de Importância dos atributos definido pelas métricas <i>Classifier Attribute Evaluation (ClassAE)</i> , <i>Gain Ratio Attribute Evaluation (GRAE)</i> e <i>Correlation Attribute Evaluation (CorrAE)</i> para a imagem de janeiro/2019.....	75
Tabela 10 - <i>Ranking</i> de Importância dos atributos definido pelas métricas <i>Classifier Attribute Evaluation (ClassAE)</i> , <i>Gain Ratio Attribute Evaluation (GRAE)</i> e <i>Correlation Attribute Evaluation (CorrAE)</i> para a imagem de abril/2019	77
Tabela 11 - <i>Ranking</i> de Importância dos atributos definido pelas métricas <i>Classifier Attribute Evaluation (ClassAE)</i> , <i>Gain Ratio Attribute Evaluation (GRAE)</i> e <i>Correlation Attribute Evaluation (CorrAE)</i> para a imagem de julho/2018	78
Tabela 12 – Composição dos conjuntos de atributos definidos para classificação	81
Tabela 13 – Número de nós das camadas escondidas da RNA e valores do parâmetro <i>mtry</i> do FA, referentes aos cinco conjuntos de atributos.....	82
Tabela 14 – Valores de RMSE obtidos a partir das classificações dos dados de validação (20% da amostra) para os diferentes conjuntos de atributos e datas de aquisição das imagens	83
Tabela 15 – Indicadores estatísticos da validação e tempo de processamento do treinamento dos classificadores FA e RNA referentes ao conjunto de atributos 5	84
Tabela 16 – Valores de <i>F-Score</i> gerados no treinamento dos classificadores FA e RNA referentes ao conjunto de atributos 5.....	84
Tabela 17 – Proporção de pixels com maior incerteza na atribuição da classe para os diferentes métodos de classificação e datas de tomada das imagens	98
Tabela 18 – Proporção de pixels classificados para cada classe para os diferentes métodos de classificação e datas de tomada das imagens	98

1. INTRODUÇÃO

O Brasil é o maior produtor de café do mundo com uma safra superior a 61 milhões de sacas beneficiadas em 2018, ano de bienalidade positiva (CONAB, 2018). Minas Gerais é o estado com maior produção de café do país, produzindo sozinho mais de um milhão de sacas que o segundo maior produtor mundial, o Vietnã (ICO, 2019).

Após um intensivo processo de modernização no seu cultivo, o café do Cerrado Mineiro alcançou uma produtividade expressiva e alta qualidade do produto, o que se refletiu na safra de 2018 que superou em mais de 90% a do ano anterior (CONAB, 2018). Como em qualquer cultura agrícola, alguns fatores prejudicam o processo de crescimento e de produção de cafeeiros, destacando-se os nematoides, microorganismos do solo, capazes de parasitar o sistema radicular da planta durante praticamente todo o ciclo da cultura (EPAMIG, 2011). Esse microorganismo tem baixa mobilidade e impede que a planta absorva os nutrientes necessários e que tenha uma vida saudável por um período longo de tempo, permanecendo após a colheita e afetando as próximas safras.

As infestações por nematoides são mais comuns em regiões de solos arenosos, bem como em terras já degradadas. Solos arenosos apresentam grande variação em suas características morfológicas e físicas sendo, geralmente, pobres em nutrientes, muito intemperizados, com elevada acidez e baixo teor de matéria orgânica (MALAVOLTA e KLIEMANN, 1985). Plantas atacadas por nematoides caracterizam-se pelo baixo vigor e comprometimento no desenvolvimento de sua parte aérea. A detecção de plantas infectadas por nematoides é dispendiosa e demanda a coleta e análise laboratorial de um grande número de amostras de solos infectados sendo, muitas vezes, uma abordagem inviável para o produtor.

O sensoriamento remoto permite a aquisição de informações sobre objetos ou fenômenos remotamente. O aprimoramento contínuo dos sistemas orbitais e suas características vêm disseminando cada vez mais sua utilização em diversas aplicações, contribuindo significativamente no estudo dedicado a diagnósticos agrícolas. O aumento das resoluções espacial e temporal, além da inclusão de bandas espectrais diferenciadas, potencializa sua utilização em estudos agrícolas. A aplicação do sensoriamento remoto no estudo da vegetação permite investigar e caracterizar o processo de interação entre a energia eletromagnética e a planta possibilitando, por exemplo, a detecção e monitoramento de doenças devido à variação espectral que a doença causa na folha. Nesse sentido, as imagens multiespectrais podem ser utilizadas para extrair informações com

base no comportamento espectral da planta e da cultura. Barnes (2000), afirma que as respostas das plantas à deficiência de nutrientes podem ser detectadas usando técnicas de sensoriamento remoto.

A Agência Espacial Europeia (ESA - *European Space Agency*), em parceria com o Programa GMES (*Global Monitoring for Environment and Security*), fornece informações visando melhorar a gestão do meio ambiente, entender e mitigar os efeitos das mudanças climáticas e garantir a segurança civil por meio das missões Sentinel. O sensor *MultiSpectral Instrument* (MSI) do Sentinel-2 foi concebido para produzir imagens de alta resolução geométrica e espectral. Esse sensor possui 13 bandas, sendo três bandas no espectro do visível e uma no infravermelho próximo (NIR – *near infrared*) com resolução espacial de 10 metros, 4 bandas no vermelho limítrofe (*Red Edge*) e 2 bandas no infravermelho de ondas curtas (SWIR - *short wave infrared*) com resolução espacial de 20 metros e ainda 3 bandas usadas na correção atmosférica da cena, com resolução de 60 metros (ESA, 2019). Sua configuração espectral, associada à possibilidade de revisita de 10 ou 5 dias (considerando 1 ou 2 satélites), tornam essas imagens promissoras nas atividades de monitoramento agrícola.

Os avanços nos sistemas sensores, assim como nos sistemas computacionais, têm favorecido a detecção precoce de doenças em culturas agrícolas. Behmann et al. (2015) destacam que a resposta registrada para uma planta possui, inerente a ela, informações advindas das condições ambientais, atmosféricas e a maneira como as medidas foram tomadas. Portanto, o estudo de doenças em culturas agrícolas, requer além de imagens corrigidas de efeitos externos, métodos avançados e robustos de análise de imagens e que permitam a integração de dados de natureza não espectral.

Behmann et al. (2015) ainda ressaltam a eficiência de métodos baseados em aprendizado de máquina na classificação de dados obtidos por sensores ópticos, já que esses métodos permitem a aquisição de dados de treinamento representativos das classes de interesse, o que pode compensar a presença de ruídos incluídos no sinal registrado.

Além disso, os algoritmos baseados em aprendizado de máquina, capazes de reconhecer padrões a partir de conjuntos de amostras de treinamento e generalizar a informação extraída para toda a imagem (LARY, 2016), podem incluir dados não espectrais no seu processamento e, de acordo com Vapnik (1999), tem por finalidade a construção de limites de decisão que possibilitem a maior separação entre os objetos, por meio da minimização dos erros. Alguns desses algoritmos comumente utilizados são: redes neurais artificiais (*Artificial Neural Networks* - ANN), máquinas de vetores de

suporte (*Support Vector Machines* - SVM), árvores de decisão (*Decision Trees* - DT) e florestas aleatórias (*Random Forest* – RF).

Independente das características das informações a serem extraídas e do método de classificação a ser adotado, alguns procedimentos são essenciais para um bom resultado, entre eles, a seleção de um conjunto de atributos que possa discriminar as informações de interesse (NOVO, 2010). Na classificação, selecionar subconjuntos de atributos é um processo para identificar e remover o máximo possível de dados irrelevantes e redundantes. Isso reduz a dimensionalidade dos dados e pode permitir que os algoritmos de aprendizado operem com mais rapidez e eficácia. Em alguns casos, a precisão da classificação pode ser melhorada; em outros, o resultado é uma representação mais compacta e pode ser mais facilmente interpretada (HALL, 1999).

Martins et al. (2017) desenvolveram uma pesquisa para a detecção de nematoides em plantações de café com base em imagens multiespectrais do sistema RadipEye. Os resultados mostraram que os dados multiespectrais indicaram eficientemente as plantas de café sadio em dois níveis de infecção. O sistema MSI/Sentinel-2, apesar de adquirir imagens em escala menor, possui uma configuração espectral que inclui duas bandas espectrais no SWIR e quatro bandas espectrais no Red-Edge, o que não ocorre com o sistema RapidEye.

Nesse contexto, a hipótese desta pesquisa é que a inclusão de bandas do SWIR do sistema Sentinel-2, trará maior eficiência na detecção de nematoides em talhões de café, pois essa região do espectro é sensível ao teor de umidade da vegetação. Com sua utilização espera-se maior atingir maior precisão na identificação de áreas atingidas, devido ao comprometimento do sistema radicular das plantas infectadas, tornando ineficiente a absorção de água e minerais causada pelo parasita (Oliveira e Rosa, 2018). Além disso, a disponibilidade de outras bandas espectrais no Red-Edge permite explorar com maior eficácia os aspectos relacionados ao vigor da vegetação, o que pode contribuir para aumentar a discriminação espectral entre áreas de café infectado e sadio.

Adicionalmente, visando aumentar o poder de detecção de doenças em plantações de café, serão utilizados algoritmos baseados em aprendizado de máquina (*machine learning*) disponíveis no sistema WEKA (*Waikato Environment for Knowledge Analysis*).

2. OBJETIVOS

2.1 Objetivo geral

Investigar o potencial das imagens multiespectrais do MSI/Sentinel-2 na detecção de áreas parasitadas por nematoides em plantação cafeeira, assim como desenvolver um método para mapeamento de áreas infectadas pelo parasita baseado em sistema de aprendizado de máquina e no ciclo de desenvolvimento da cultura.

2.2 Objetivos específicos

Os objetivos específicos da pesquisa são:

- Avaliar o potencial de índices espectrais que considerem as bandas do SWIR (infravermelho de ondas curtas) na detecção de áreas infectadas por nematoides;
- Avaliar a importância dos atributos que melhor discriminem entre talhões de café infectados por nematoides e café saudável;
- Verificar a capacidade de algoritmos baseados em aprendizado de máquina no mapeamento de talhões de café infectados por nematoides;
- Mapear as áreas de ocorrência de nematoides na cultura cafeeira e estimar a qualidade de mapas temáticos produzidos, considerando as fases fenológicas representativas da cultura.

3. JUSTIFICATIVA

Diversas são as possibilidades de aplicação de técnicas de sensoriamento remoto no setor agrícola. O avanço tecnológico e desenvolvimento de novos sensores, plataformas de aquisição de dados e ferramentas computacionais, permitem manipular grandes conjuntos de dados, sendo absorvidos em grande escala na agricultura.

O uso de dados de sensoriamento remoto na agricultura vem sendo um assunto cada vez mais abordado na literatura. Na edição especial de dezembro/2017 do periódico *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (IEEE-JSTARS), intitulada *Special Issue on Agro-Geoinformatics: Monitoring, Prediction and Decision Support*, Busetto et al. (2017) cita a importância da disponibilidade de informações de alta qualidade espacial e temporal das culturas, assim como a condição dessas culturas, para o monitoramento e melhoria de produtividade, apoiando soluções em escalas local e continental e minimizando os impactos ambientais. Autores como Zhang et al. (2017) e Gonzalo-Martin et al. (2017) mostram a utilização do classificador Florestas Aleatórias (*Random Forest*) para classificação e mapeamento de culturas agrícolas, a partir de imagens RapidEye e Landsat-7, respectivamente. Ainda nessa edição especial, Sun et al. (2017) avaliaram diferentes índices de vegetação no monitoramento da seca em áreas agrícolas.

Quando o cafeeiro está infectado por nematoides, a resposta espectral da folha sofre variações devido a desequilíbrios nutricionais na planta e com a disseminação da doença, a área infectada apresenta uma configuração espacial característica, passível de detecção em imagens multiespectrais (MARTINS et al., 2017). Considerando a importância da detecção de doenças em plantações de café, o uso do sensoriamento remoto aliado a outras tecnologias, tem sido de grande relevância para evitar maiores prejuízos em plantações cafeeiras, possibilitando o tratamento da área após a detecção e também seu monitoramento para futuros problemas.

Diante de diversos estudos bem-sucedidos e continuados utilizando métodos baseados em aprendizado de máquina, julga-se oportuno e relevante explorar seu potencial no monitoramento de culturas agrícolas a partir da utilização de imagens multiespectrais de média resolução espacial, face à vasta disponibilidade desses dados e ampla possibilidade de produzir imagens derivadas das bandas espectrais.

É fato que a diminuição da produtividade provocada por pragas, doenças ou escassez de nutrientes e água devem ser diagnosticadas de forma ágil a fim de fornecer

informações fundamentais para aumentar a eficiência do manejo e alcançar as expectativas de produção agrícola. Nesse contexto, a tecnologia e novas metodologias podem contribuir para otimizar sistemas de produção agrícola e adicionalmente diminuir o impacto ambiental desses processos de produção.

Assim, este trabalho contribui com a comunidade científica tratando de um tema que é relevante e atual no contexto da agricultura de precisão, além do desenvolvimento de uma metodologia que subsidie a detecção de nematoides em produções cafeeiras.

4. FUNDAMENTAÇÃO TEÓRICA

4.1 Fenologia do cafeeiro

O ciclo fenológico do cafeeiro arábica (*Coffea arabica* L.) apresenta uma sucessão de fases vegetativas e reprodutivas, que ocorrem em aproximadamente 2 anos, diferente daquele da maioria das plantas, que emitem as inflorescências na primavera e frutificam no mesmo ano fenológico (CAMARGO e CAMARGO, 2001). Um modelo de esquematização da sequência dos estágios fenológicos do cafeeiro arábica foi apresentado pelos autores, com base nas condições climáticas tropicais do Brasil (Figura 1).

O ciclo fenológico, para as condições tropicais do Brasil, foi subdividido em seis fases distintas, sendo que duas ocorrem no período vegetativo (1º ano fenológico), e outras quatro no reprodutivo (2º ano fenológico).

Figura 1 - Esquematização das seis fases fenológicas do cafeeiro arábica, durante 24 meses

1º ano fenológico											
Set.	Out.	Nov.	Dez.	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.
(1) Vegetação e formação das gemas foliares							(2) Indução e maturação das gemas florais				
Período Vegetativo										Repouso	

2º ano fenológico											
Set.	Out.	Nov.	Dez.	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.
(3) Florada, chumbinho e expansão dos frutos				(4) Granação dos frutos			(5) Maturação dos frutos			(6) Repouso e senescência dos ramos terciários e quaternários	
Período Reprodutivo										Autopoda	
Novo período vegetativo											

Fonte: Adaptado de Camargo e Camargo (2001)

Camargo e Camargo (2001) explicam as fases dos dois anos fenológicos do cafeeiro. A primeira fase, apresentada como (1) vegetação e formação das gemas foliares, ocorre normalmente de setembro a março, meses de dias longos, com fotoperíodo de 13 a 14 horas de luz efetiva ou acima de 12 horas de brilho solar.

A segunda fase, (2) indução e maturação das gemas florais, acontece normalmente de abril a agosto, meses de dias mais curtos. A partir de fevereiro, quando

os dias começam a ficar mais curtos, intensifica-se o crescimento das gemas florais, as quais, após completo o seu desenvolvimento, entram em dormência e ficam prontas para a antese, quando acontece um aumento substancial de seu potencial hídrico, causado por chuva ou irrigação. Nos dois meses finais (julho e agosto), as gemas dormentes produzem um par de folhas pequenas, etapa essa que separa o primeiro ano fenológico do segundo.

A fase (3) florada, chumbinho e expansão dos frutos, que vai de setembro a dezembro, inicia-se com a florada dos frutos, depois de um aumento do potencial hídrico nas gemas florais maduras (choque hídrico), e vai até a expansão dos frutos. Uma florada principal acontece quando se verifica um período de restrição hídrica seguido de chuva, irrigação ou mesmo de um acentuado aumento da umidade relativa do ar. A ocorrência de uma forte estiagem nessa fase, pode prejudicar o crescimento dos frutos.

A quarta fase é a de granação dos frutos (4), que ocorre em pleno verão, de janeiro a março do segundo ano fenológico (período reprodutivo), quando há formação dos grãos. Nessa fase, um estresse hídrico pode ser prejudicial, produzindo frutos mal granados. A maturação dos frutos se dá na quinta fase (5), compreendendo normalmente os meses de abril, maio e junho. Nessa etapa, a evapotranspiração potencial (ET_p) decresce significativamente e as deficiências hídricas moderadas beneficiam a qualidade do produto. A sexta e última fase (6), de julho e agosto, constitui a de senescência dos ramos produtivos não-primários (terciários e quaternários), que secam e morrem, condicionados à conhecida autopoda dos cafeeiros.

4.2 Nematoides na plantação de café

O café é o principal produto de exportações do agronegócio de Minas Gerais, sendo considerado um importante gerador de emprego e renda para as famílias mineiras (EMATER, 2018). No entanto, distúrbios de naturezas diversas podem acometer os cafeeiros, provocando anormalidades que podem impedir o seu crescimento bem-sucedido, levando à baixa produtividade. Dentre os fatores limitantes ao processo de crescimento e de produção do cafeeiro, destacam-se os nematoides.

Os nematoides são organismos multicelulares podendo ser classificados como parasitas de vida livre ou até mesmo parasitas de animais ou plantas (IBRAHIM et al., 2019). São diversas as formas de interação dos nematoides parasitas de plantas com seus hospedeiros. Alguns nematoides são ectoparasitos migratórios, ou seja, não entram no

hospedeiro, mas migram através do solo, usando raízes como fonte de alimento. Os endoparasitas migratórios entram no hospedeiro e migram através dos seus tecidos, permanecendo grande parte do seu ciclo de vida dentro das raízes das plantas. Há também os nematoides semi-endoparasitários, que podem ter tanto estágios migratórios, quanto penetrar na planta hospedeira para se alimentar em algum estágio do seu ciclo de vida. Considerando seu modo de parasitismo, os nematoides podem não apenas retirar da célula vegetal os nutrientes essenciais para seu desenvolvimento e reprodução, mas também viabilizar a injeção de substâncias tóxicas no interior da mesma (JONES et al., 2013).

Ainda de acordo com o modo de parasitismo, os nematoides mais importantes para a indústria cafeeira são os sedentários (endoparasitas), ou seja, passam grande parte da sua vida se alimentando de células da raiz no mesmo local, alterando o seu metabolismo. Segundo a EPAMIG (2011), diversas espécies encontradas estão associadas às raízes de cafeeiros no Brasil, sendo as espécies dos gêneros *Meloidogyne* e *Pratylenchus* as mais prejudiciais.

A EPAMIG (2011) reporta que as espécies dos nematoides do gênero *Meloidogyne* são as mais danosas para a cafeicultura mundial e também as mais disseminadas em culturas cafeeiras no Brasil, sendo a espécie *M. exigua* a mais encontrada (Rio de Janeiro, São Paulo, Minas Gerais, Paraná, Espírito Santo, Bahia, Distrito Federal), seguida de *M. incógnita* (Paraná, São Paulo, Minas Gerais, Espírito Santo, Rio de Janeiro) e *M. paranaensis* (Paraná, Minas Gerais, São Paulo, Goiás).

Segundo Oliveira e Rosa (2018), o ataque dos nematoides do gênero *Meloidogyne* pode, em geral, tornar o sistema radicular da planta ineficiente na absorção de água e nutrientes. As plantas podem apresentar coloração anormal e sintomas de deficiência nutricional (Figura 2), ocasionando a diminuição na produção. Além disso, outras condições de solo e clima também podem intensificar esses sintomas. Ainda de acordo com Oliveira e Rosa (2018), as áreas mais afetadas pela presença desses parasitos são as de solos arenosos e degradados com baixa fertilidade e deficiência hídrica.

Figura 2 - Lavoura nova de café arábica implantada em área infestada por *M. paranaensis*



Fonte: EPAMIG, 2011.

Devido à sua baixa capacidade de deslocamento, a introdução e disseminação dos nematoides ocorre principalmente pelo solo, através da movimentação de máquinas e implementos agrícolas contaminados, ou através de erosão hídrica ou eólica (IAC, 2007).

Em terras brasileiras, aproximadamente 20% do que é produzido pelo cafeicultor não chega ao mercado, situação que poderia ser revertida adotando medidas de controle ou até mesmo ações de caráter preventivo, sendo mais eficientes e econômicas quando comparadas aos tratamentos curativos (OLIVEIRA e ROSA, 2018). O controle dos nematoides é, em geral, difícil de ser realizado, considerando que, em área já contaminada, sua erradicação é praticamente impossível. Entretanto, esses parasitos podem ter suas populações reduzidas e mantidas em níveis baixos através da adoção de medidas de controle (IAC, 2007).

Nesse contexto, a prevenção é a melhor forma de controle desses parasitos, sendo essencial a adoção de práticas de contenção da disseminação dos nematoides em áreas ainda não infectadas. A contenção de enxurradas, a limpeza de todos os meios de transporte e equipamentos agrícolas que circulam nas propriedades, a escolha da área para implantação da lavoura cafeeira (área sem nenhuma plantação por vários anos, por exemplo), a utilização de cultivares resistentes, o controle químico como o uso de nematicidas e a adição de matéria orgânica (reduz a população de nematoides, e podem

favorecer a estruturação física e química do solo), constituem de medidas de manejo de extremo valor que impedem a introdução e disseminação dos nematoides (EPAMIG, 2011).

Uma pesquisa desenvolvida por Martins et al. (2017), teve como objetivo a detecção e o mapeamento de infecção de nematoides em plantações de café de uma área cafeeira localizada no sul do estado de Minas Gerais, na qual foi constatada a presença de nematoides. Para isso, foram utilizados dados biofísicos da cultura, medições radiométricas obtidas *in situ* e imagens multiespectrais do sistema RapidEye. Além disso, foram identificadas as bandas espectrais do sensor mais sensíveis para a detecção desse problema no café. Seus resultados mostraram que os dados biofísicos não representaram de forma fiel a distribuição espacial do parasita, mas os dados multiespectrais, mais especificamente as bandas do vermelho, do vermelho limítrofe (*Red Edge*) e do infravermelho próximo, reproduziram mais eficientemente, as áreas de café sadio em dois níveis de infecção, apresentando acurácia global de 78% e coeficiente de *Kappa* de 0,71, o que para o assunto abordado, foi considerado um resultado adequado.

Um estudo desenvolvido por Mutanga et al. (2017) aponta a necessidade de se detectar, mapear e monitorar doenças para garantir a segurança alimentar, em particular na África subsaariana, onde a fome e a pobreza atingiram estágios alarmantes. Os autores afirmam que a infecção de culturas por doenças ou efeitos das mudanças climáticas podem resultar em perdas de rendimentos significativas. Assim, o estudo teve por objetivo realizar uma pesquisa bibliográfica para avaliar o potencial de dados de sensoriamento remoto de alta resolução na detecção de áreas de cultivos infectadas, bem como áreas vulneráveis a esse problema. A revisão realizada pelos autores mostrou que houve um aumento significativo na aplicação de técnicas de sensoriamento remoto na detecção de cultivos infectados e que o sensoriamento remoto proporcionou a oportunidade de detectar e mapear culturas afetadas por doenças e pragas mais eficientemente. Porém, essa abordagem ainda é um desafio, principalmente em economias pobres, devido ao trabalho envolvido e custos de aquisições dos dados.

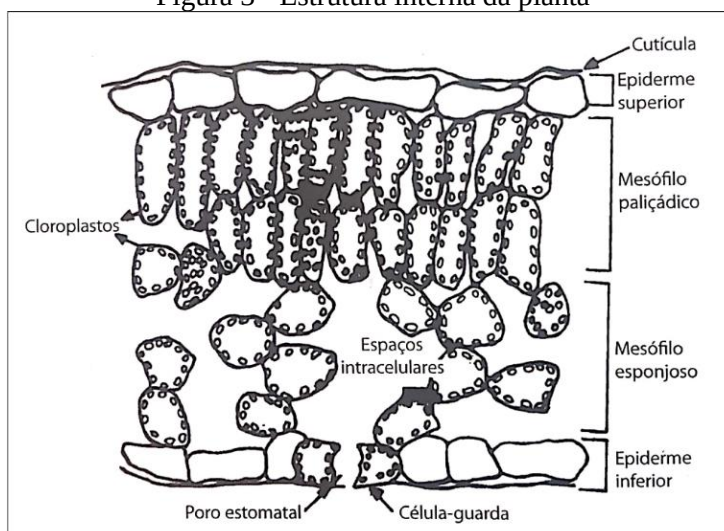
4.3 Sensoriamento remoto da vegetação

4.3.1 Interação da radiação eletromagnética com a vegetação

De acordo com Jensen (2009), a maior parte da superfície terrestre é coberta por diversos tipos de vegetação o que levou ao grande empenho no desenvolvimento de tecnologias capazes de inferir informações biofísicas da vegetação a partir de dados de sensoriamento remoto. Segundo Ponzoni (2012), o mapeamento de classes específicas da cobertura vegetal brasileira por meio de imagens de sensoriamento remoto encontra-se em pleno desenvolvimento, podendo, por exemplo, prever safras e oferecer suporte à preservação do meio ambiente.

Para entender o processo de interação da radiação eletromagnética com a planta, é importante conhecer sua estrutura interna (Figura 3). Diferentes espécies de plantas possuem variações nas suas estruturas internas, porém, elas são estruturas pelas quais a radiação eletromagnética trafega e dependendo do comprimento de onda dessa radiação, a interação se comporta de maneira diferente.

Figura 3 - Estrutura interna da planta



Fonte: Ponzoni, 2012

Utilizando a Figura 3, que se refere a um corte transversal em uma determinada folha, Ponzoni (2012) explica essa estrutura. A superfície da folha (face ventral) é a parte que mais recebe radiação eletromagnética proveniente do Sol exercendo o papel de proteção com algumas estruturas como pelos e camadas de cera (cutícula). A epiderme, camada abaixo, é composta geralmente por células alongadas, entre outras, com o papel

de formar estômatos, que ajudam a planta a manter o balanço entre a perda de água e suas necessidades de oxigênio e gás carbônico, o qual depende de condições ambientais e fisiológicas. O mesófilo paliçádico é organizado por células ricas em cloroplastos onde, internamente, encontram-se os pigmentos fotossintetizantes: os carotenoides, as ficobilinas e principalmente as clorofilas. Logo abaixo, na camada do mesófilo esponjoso, sua organização de células é menos compacta em relação à camada superior, apresentando maior espaço entre as células, o qual é preenchido por gases resultantes dos processos de respiração e de transpiração. Por fim, a epiderme inferior apresenta frequentemente menor número de estômatos, e também é constituída por pelos e ceras.

O processo de interação da radiação eletromagnética (REM) com os diversos alvos da superfície terrestre é descrito pelos fenômenos da reflexão, transmissão e absorção, os quais dependem de fatores como o comprimento de onda, ângulo em que a energia incidente atinge a superfície e a composição dos materiais da superfície. São essas diferenças que tornam possível distinguir entre as múltiplas feições da terra visíveis nas imagens orbitais (LINDGREN, 1985).

Na vegetação, a magnitude das porções de REM absorvidas, transmitidas e refletidas pela planta, torna capaz diferir entre diversas espécies ou até mesmo eventos entre a mesma espécie. Dos pontos de vista fisiológico e bioquímico, a porção de energia absorvida pela planta é a mais importante. Os pigmentos contidos na folha são responsáveis por aproximadamente 50% da absorção de REM que chega até a planta. Essa energia absorvida é responsável por alguns fenômenos como a fotossíntese e fotoconversão. Dentre as funções principais da planta, a fotossíntese é a única diretamente ligada à radiação solar. No processo de fotossíntese, é absorvida a radiação solar de comprimentos de onda entre 400 e 700 nm, ou seja, a energia utilizada pela planta corresponde à luz azul (400-530 nm) e à vermelha (600-730 nm), as quais são absorvidas pelos pigmentos de cloroplastos (MOREIRA, 2012).

Ainda segundo Moreira (2012), os componentes principais dos cloroplastos são os pigmentos e as clorofilas. Nas plantas superiores (plantas que produzem flores, sementes e frutos) são encontrados os pigmentos clorofila “a” (máxima absorção da luz de 430 a 660 nm), clorofila “b” (de 453 a 643 nm), “ α -Caroteno” (em 420, 440 e 470 nm), “ β -Caroteno” (em 425, 450 e 480 nm) e xantofilas: Lutol (em 425, 445 e 475 nm) e Violoxanto (em 425, 450 e 475 nm).

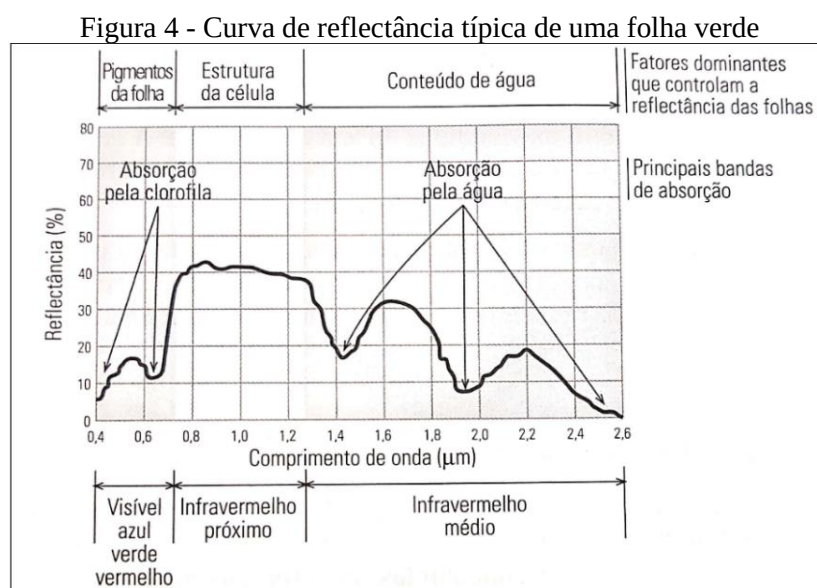
A medida de reflectância de uma cobertura vegetal é consideravelmente menor do que a de uma simples folha devido à atenuação da radiação pela variação de ângulos de

incidência da radiação e de visada. No entanto, existe uma grande relação entre elas pois há predominância de folhas no dossel da vegetação em relação à outras partes da planta (JENSEN, 2009).

Os principais elementos responsáveis pela reflectância da planta são: pigmentos, espaços ocupados pela água e pelo ar e estruturas celulares que possuem mesma dimensão do comprimento de onda da radiação incidente. As células da camada superficial (cutícula ou epiderme) são responsáveis pela reflectância de pequena parte da radiação incidente. A maior parte dessa energia é transmitida para o mesófilo esponjoso, onde os raios interagem com o meio e são refletidos. Dado ao grande número de paredes celulares internas à planta, alguns raios são refletidos em direção à fonte de energia e outros são transmitidos completamente através da planta (MOREIRA, 2012).

4.3.2 Comportamento Espectral da Vegetação

Conhecer o comportamento espectral dos alvos é de fundamental importância para a extração de informações por meio de imagens orbitais. O comportamento espectral da vegetação é determinado por diversos fatores, incluindo a sua composição, morfologia e estrutura interna (PONZONI, 2002). Além disso, as características de reflectância de alvos vizinhos também podem influenciar na resposta do objeto de interesse, podendo mascarar a sua resposta. A curva de reflectância de uma folha verde é mostrada na Figura 4.



Fonte: Novo (2010)

A Figura 4 pode ser decomposta em três principais regiões espectrais: região do visível (400 a 700 nm); região do infravermelho próximo (700 a 1300 nm); e região do infravermelho médio (1300 a 2600 nm) (NOVO, 2010).

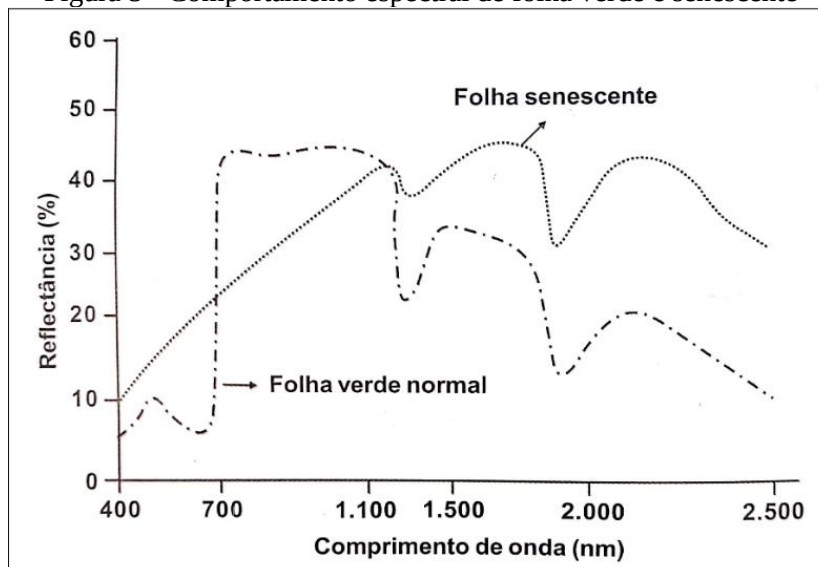
Na região do visível, a maior parte da energia incidente é absorvida pelos pigmentos fotossintetizantes presentes no mesófilo da planta, que são os pigmentos compostos principalmente pelas clorofilas “a” e “b”, e em menor quantidade pelos carotenoides, xantofilas e antocianinas. Nessa região, tanto a reflectância quanto a transmitância das folhas são menores que 20%, no entanto, há um pequeno aumento na reflectância da planta no comprimento de onda de 555 nm, o qual é responsável pela percepção do olho humano da cor verde na vegetação (NOVO, 2010).

Na região do infravermelho próximo, percebe-se uma alta reflectância da vegetação, variando de 30 a 40% e uma baixa absorção (menos de 15%). Isso se deve à interação da radiação eletromagnética com a estrutura celular interna da planta (mesófilo), ou seja, quanto mais espaço na estrutura interna da planta, maior o espalhamento interno da radiação incidente, e assim, maior será a reflectância (PONZONI, 2002).

Já na região do infravermelho médio, a reflectância da planta caracteriza-se pelo decréscimo dos valores de energia refletida, em razão, principalmente, do conteúdo de água presente na folha. A radiação eletromagnética é absorvida consideravelmente na região entre 1300 nm e 2000 nm, onde se encontram os dois pontos máximos de absorção pela água, em 1400 nm e em 1950 nm (NOVO, 2010).

Existem diversos fatores que interferem na reflectância da folha, como o conteúdo de água, a idade da folha, posição nodal, condição de iluminação, pubescência e senescência. A idade da planta, por exemplo, interfere diretamente na capacidade da planta realizar a fotossíntese, processo intimamente relacionado à quantidade de REM absorvida nas faixas do azul e do vermelho. A planta senescente manifesta maior reflectância na região do visível do que a planta verde, em razão da degradação dos pigmentos (clorofilas) que são responsáveis pela absorção da energia (MOREIRA, 2012). As diferenças na reflectância espectral entre uma folha verde e uma folha senescente são mostradas na Figura 5.

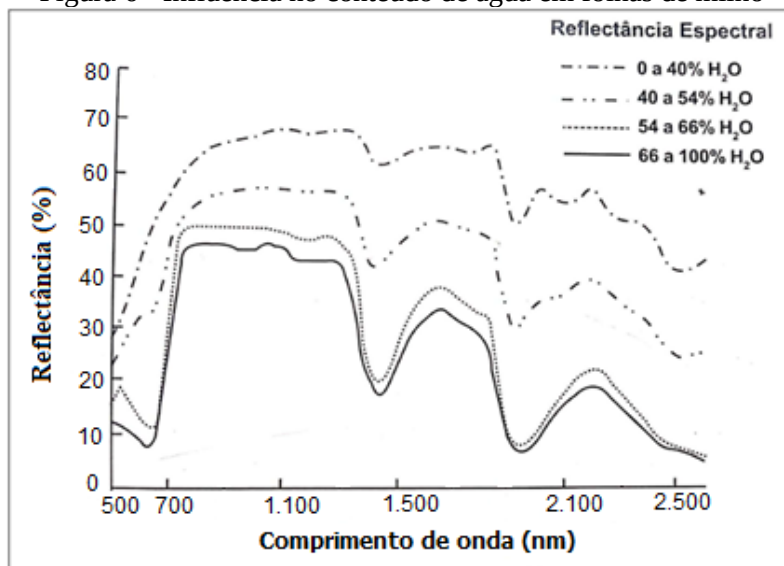
Figura 5 - Comportamento espectral de folha verde e senescente



Fonte: Moreira (2012)

O conteúdo de água na planta também interfere no seu comportamento espectral, como pode ser visto na Figura 6. O déficit de água na folha causa diversas reações internas, sendo a primeira, o fechamento dos estômatos, fazendo com que a planta diminua a perda de água e a sua capacidade fotossintética (MOREIRA, 2012), aumentando a sua reflectância ao longo do espectro.

Figura 6 - Influência no conteúdo de água em folhas de milho



Fonte: Moreira (2012)

Em geral, na região do visível, as plantas com menor teor de água possuem menor absorção da REM em relação às plantas com maior teor de água, devido à redução do conteúdo de clorofila da planta, aumentando, consequentemente sua reflectância. As

mudanças internas na estrutura da folha, também aumentam a reflectância na região do infravermelho próximo (MOREIRA, 2012).

4.4 Índices espectrais

O sensoriamento remoto tem sido amplamente utilizado no estudo da cobertura vegetal da superfície terrestre e suas particularidades, pois permite obter grande número de informação a partir de imagens multiespectrais facilitando a análise e a interpretação dos dados de superfície. Os índices espectrais gerados a partir de imagens auxiliam no monitoramento de alterações da cobertura terrestre, decorrentes tanto de eventos naturais quanto por eventos de natureza antrópica.

Os índices de vegetação foram criados com o objetivo de realçar o contraste espectral entre a vegetação e os alvos vizinhos. Eles se tornaram grandes aliados no mapeamento da cobertura vegetal, capazes de realizar o monitoramento em culturas agrícolas, melhorar a produtividade e detectar doenças, como pode ser visto em Busseto et al.; Zhang et al.; Gonzalo-Martín et al. e Martins et al. (2017). Eles são obtidos por meio de operações matemáticas entre valores numéricos de pixels de bandas espectrais específicas, com o objetivo de destacar algum evento de interesse.

Diversos índices de vegetação foram propostos ao longo do tempo e estão disponíveis na literatura, dentre eles podem ser citados: Razão Simples (RS), *Normalized Difference Vegetation Index* (NDVI), *Normalized Difference Water Index* (NDWI), *Soil-Adjusted Vegetation Index* (SAVI), *Enhanced Vegetation Index* (EVI), *Atmospheric Resistant Vegetation Index* (ARVI), *Vogelmann Red-Edge Index* (VOG1), *Char Soil Index* (CSI), *Variation of Normalized Burn Ratio* (NBR2), *Modified Chlorophyll Absorption in Reflectance Index* (MCARI), *Moisture Stress Index* (MSI), *Transformed Chlorophyll Absorption in Reflectance Index* (TCARI) e *Normalized Ratio Indices* (NRI). Buscando uma maior discriminação entre o café infectado e o café sadio, são utilizados, além do clássico NDVI, aqueles cuja formulação considere as bandas do infravermelho de ondas curtas (SWIR) e do Red-Edge (RE) (Tabela 1).

Tabela 1 – Formulação dos índices de vegetação utilizados. NIR (*Near Infrared*) se refere ao infravermelho próximo (comprimento de onda central 832,8 nm), RED ao vermelho (664,6 nm), SWIR1 (1613,7 nm) e SWIR2 (2202,4 nm) aos comprimentos de onda do infravermelho de ondas curtas 1 e 2 (*Short Wave Infrared*), RE1 ao *Red-Edge* 1 (704,1 nm) e GREEN ao canal verde (559,8 nm)

Índices	Equações	Referência
NDVI	$\frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}}$	Rouse et al. (1974)
NDWI	$\frac{\rho_{NIR} - \rho_{SWIR1}}{\rho_{NIR} + \rho_{SWIR1}}$	Gao (1996)
CSI	$\frac{\rho_{NIR}}{\rho_{SWIR2}}$	Smith et al. (2005)
NBR2	$\frac{\rho_{SWIR1} - \rho_{SWIR2}}{\rho_{SWIR1} + \rho_{SWIR2}}$	Key e Benson (1999)
MSI	$\frac{\rho_{SWIR1}}{\rho_{NIR}}$	Hunt e Rock (1989)
MCARI	$[(\rho_{RE1} - \rho_{RED}) - 0,2(\rho_{RE1} - \rho_{GREEN})] \left(\frac{\rho_{RE1}}{\rho_{RED}} \right)$	Daughtry et al. (2000)
TCARI	$3[(\rho_{RE1} - \rho_{RED}) - 0,2(\rho_{RE1} - \rho_{GREEN})] \left(\frac{\rho_{RE1}}{\rho_{RED}} \right)$	Haboudane et al. (2002)
NRI	$\frac{\rho_{SWIR1} - \rho_{RED}}{\rho_{SWIR1} + \rho_{RED}}$	Herrmann et al. (2010)
MCARI*	$[(\rho_{RE1} - \rho_{SWIR1}) - 0,2(\rho_{RE1} - \rho_{GREEN})] \left(\frac{\rho_{RE1}}{\rho_{SWIR1}} \right)$	Herrmann et al. (2010), modificado de Daughtry et al. (2000)
TCARI*	$3[(\rho_{RE1} - \rho_{SWIR1}) - 0,2(\rho_{RE1} - \rho_{GREEN})] \left(\frac{\rho_{RE1}}{\rho_{SWIR1}} \right)$	Herrmann et al. (2010), modificado de Haboudane et al. (2002)

O NDVI foi criado por Rouse et al. em 1974, com o objetivo de realizar uma medição quantitativa de condições de vegetação em amplas regiões usando dados Landsat-1. É uma medida radiométrica adimensional que indica o nível de cobertura vegetal presente na superfície terrestre e é uma simples combinação das refletâncias do vermelho e do infravermelho próximo, considerando que a banda do vermelho está localizada na região de alta absorção da planta, enquanto a do infravermelho próximo está localizada na região de alta refletância de copas de vegetação.

O Índice de Água por Diferença Normalizada ou Índice de Umidade por Diferença Normalizada (NDWI), proposto por Gao (1996), é outro importante índice no estudo da cobertura vegetal, permitindo ressaltar feições de água e minimizar o restante dos alvos. Esse índice é sensível às variações no conteúdo de água do dossel da vegetação, devido à utilização de uma banda espectral do SWIR.

Assim como a banda do SWIR apresenta potencial na discriminação entre áreas infectadas e não infectadas por nematoides, devido à sua sensibilidade ao conteúdo de água da vegetação, ela também pode auxiliar na detecção de outros fenômenos. Smith et al. (2007) utilizaram diversos índices espectrais para avaliar a eficácia na discriminação entre superfícies queimadas e não queimadas em uma série de escalas espaciais. A pesquisa mostrou que índices espectrais que incorporam os comprimentos de onda do SWIR (dentre eles o CSI) apresentaram um desempenho melhor do que os índices que incluem apenas as bandas NIR e bandas do visível na discriminação de áreas queimadas.

O índice NBR2 (KEY e BENSON, 1999) é uma variação do índice *Normalized Burn Ratio* (NBR) que foi criado com o intuito de destacar áreas de vegetação com diferentes conteúdos de água, além de também poder auxiliar estudos de recuperação de áreas queimadas, a partir de sua discriminação prévia.

Hunt e Rock (1989) propuseram o índice MSI com o objetivo de realizar a detecção de plantas submetidas a estresse hídrico por meio de sensoriamento remoto, utilizando comprimentos de onda do infravermelho próximo e infravermelho de ondas curtas (SWIR).

Herrmann et al. (2010), desenvolveram uma pesquisa com o objetivo de explorar novos índices espectrais dependentes da região do infravermelho de ondas curtas (1200-2500 nm), mas principalmente a banda de 1510 nm, pois está diretamente relacionada ao conteúdo de nitrogênio na planta. Como os índices MCARI e TCARI são índices sensíveis à presença da clorofila, porém sem ligação direta com o conteúdo de nitrogênio, os autores substituíram a banda do vermelho pela banda do SWIR nesses índices (MCARI* e TCARI*). Os mesmos autores ainda propuseram o NRI que consiste em relacionar a característica de absorção de clorofila na região do vermelho (660 nm) e na absorção de nitrogênio na região do SWIR (1510 nm). Assim como em Herrmann et al. (2010), os índices MCARI e TCARI são utilizados nessa pesquisa, com base na banda do infravermelho de ondas curtas.

4.5 Imagem de textura

Quando os seres humanos interpretam visualmente as imagens obtidas remotamente, geralmente levam em consideração o contexto, as bordas, a textura e a variação de tonalidade. Por outro lado, muitos dos algoritmos de classificação de imagens multiespectrais se baseiam apenas no uso de informações espectrais. A necessidade de incorporar características de natureza espacial ao processo de classificação tem levado ao desenvolvimento de procedimentos para extrair outros atributos baseados na textura da imagem. Nesse contexto, cada novo pixel da imagem de textura passa a ter um valor que representa um atributo de textura naquele local (JENSEN, 1996).

Gonzales e Woods (2010) afirmam que, embora não haja uma definição formal de textura, essa característica fornece medidas de propriedades como suavidade, rugosidade e regularidade. Uma das técnicas utilizadas para a extração de características de texturas de imagens, é o cálculo da matriz de co-ocorrência. O objetivo principal dessa matriz é caracterizar texturas em uma imagem por meio de um conjunto de estatísticas para as ocorrências de cada nível de cinza em pixels diferentes ao longo de diferentes direções.

4.5.1 Construção da matriz de co-ocorrência

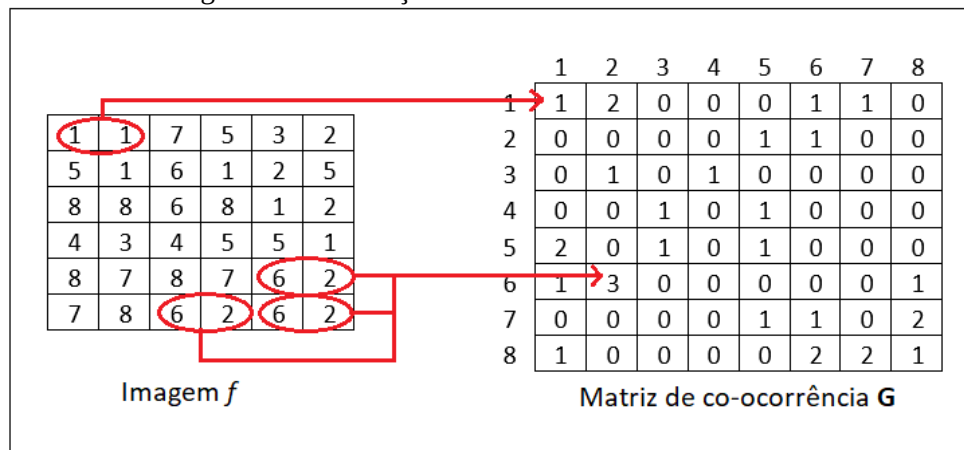
A matriz de co-ocorrência considera a relação entre dois pixels por vez, um chamado de pixel referência e o outro de pixel vizinho. O pixel vizinho pode situar-se em qualquer direção, como por exemplo, a 0° , 45° , 90° ou 135° em relação ao pixel de referência. A distância entre os dois pixels considerados, também pode variar. Cada pixel dentro da imagem torna-se o pixel referência, e a pesquisa pela direção e distância especificados se inicia no canto superior esquerdo, deslocando-se até o inferior direito. Para a construção de uma matriz de co-ocorrência com distância 1 e ângulo 0° entre dois pixels e considerando uma imagem com 8 níveis de intensidade, Gonzales e Woods (2010) descrevem os seguintes passos:

Considera-se um operador Q que define a posição de dois pixels um em relação ao outro e uma imagem f , com L níveis de intensidade possíveis. Uma imagem $\mathbf{G}$ é uma matriz cujo elemento g_{ij} é o número de vezes que os pares de pixels com intensidades z_i

e z_j ocorrem na imagem f na posição definida por Q , em que $1 \leq i, j \leq L$. Uma matriz formada dessa maneira ($\mathbf{G}$) é chamada de matriz de co-ocorrência.

A Figura 7 mostra um arranjo à esquerda como uma pequena imagem que está sendo analisada (f) e o arranjo à direita como a matriz $\mathbf{G}$ construída. Observa-se que o elemento (1,1) de $\mathbf{G}$ possui o número 1, devido à apenas uma ocorrência em f de um pixel com o valor de 1 com o outro pixel avaliado com 1, imediatamente à sua direita (distância 1 e 0°). Da mesma forma, o elemento (6,2) da matriz $\mathbf{G}$ é 3, pois existem 3 ocorrências em f de um pixel com um valor de 6 e outro pixel de 2, imediatamente à sua direita. Os demais elementos de $\mathbf{G}$ são calculados da mesma forma.

Figura 7 – Construção de uma matriz de co-ocorrência



Fonte: Gonzalez e Woods, 2010

O tamanho da matriz de co-ocorrência ($\mathbf{G}$) é definido pelo número de níveis de intensidade possíveis da imagem. Como por exemplo, para uma imagem de 16 bits (65.536 níveis de cinza), $\mathbf{G}$ será de 65.536 x 65.536.

Já que $\mathbf{G}$ depende do operador que define a posição dos dois pixels analisados (Q), a presença de padrões de textura de intensidade pode ser detectada escolhendo um operador de posição adequado e analisando elementos de $\mathbf{G}$. Alguns dos descritores utilizados para a caracterização do conteúdo de $\mathbf{G}$ são: probabilidade máxima, correlação, contraste, uniformidade, homogeneidade, entropia, entre outros. A entropia mede a aleatoriedade dos elementos de $\mathbf{G}$, e é calculada da seguinte forma:

$$Entropia = - \sum_{i=1}^k \sum_{j=1}^k p_{ij} \log_2 p_{ij} \quad (1)$$

em que p_{ij} é o i -ésimo termo de $\mathbf{G}$ dividido pela soma dos elementos de $\mathbf{G}$.

O valor da entropia é zero quando todos os p_{ij} s são 0 e é máxima quando todos os p_{ij} s são iguais. O valor máximo é $2\log_2 k$, para uma janela de tamanho $K \times K$.

4.6 Classificação baseada em aprendizado de máquina

A classificação é um processo de extração de informações a partir de imagens digitais com o objetivo de associar cada pixel (ou grupo de pixels) da imagem a uma classe temática, de acordo com o uso e ocupação do solo (MOREIRA, 2012). Em um espaço multidimensional, cada pixel da imagem é representado por coordenadas associadas ao seu nível digital (brilho), o qual depende do comportamento espectral de cada alvo, em cada banda espectral (NOVO, 2010). A distinção e identificação de cada alvo é possível devido ao comportamento característico ao longo do espectro eletromagnético dos alvos da superfície terrestre.

O processo de classificação de imagens pode ser dividido em supervisionado e não supervisionado, sendo que ambas possuem regras de decisão para que o classificador associe cada pixel à determinada classe ou região de similaridade.

Em uma classificação supervisionada, são utilizadas amostras de treinamento representativas com identidade conhecida para classificar pixels de identidade desconhecida (MOREIRA, 2012). Nessa abordagem, os níveis digitais dos pixels do conjunto dos dados de treinamento em cada uma das bandas espectrais são comparados com os níveis digitais de cada pixel da imagem, decidindo assim, à qual classe cada pixel pertence. Deve-se assegurar que os pixels escolhidos como amostras de treinamento sejam realmente representativos das classes alvo.

Os métodos de classificação de imagem podem ainda ser divididos em paramétricos e não paramétricos. Os métodos paramétricos têm caráter probabilístico e estão baseados na distribuição normal dos dados, o que demanda certa homogeneidade nas propriedades espectrais das classes. Já os classificadores não paramétricos, têm natureza determinística, ou seja, podem ser empregados em distribuições que não obedeçam aos parâmetros da curva normal (RICHARDS e JIA, 2006).

Avanços mais recentes na tecnologia de sistemas sensores aumentam consideravelmente a capacidade de coleta de um grande número de faixas espectrais de uma mesma cena. Entretanto, um aumento significativo no número de bandas, pode

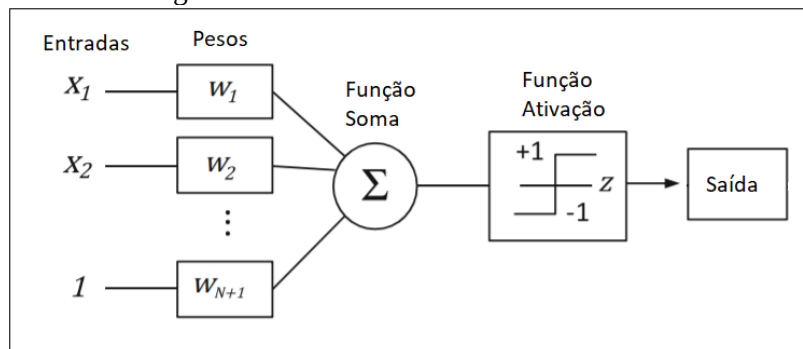
interferir na acurácia do resultado produzido por um classificador paramétrico, devido ao aumento no número de parâmetros a serem estimados, principalmente a matriz de covariância (LICZBINSKI, 2007). Nesse sentido, outras abordagens de classificação de imagens, como a aquelas realizadas por algoritmos baseados em aprendizado de máquina, são capazes de tratar dados de alta dimensionalidade sem a necessidade de pressupor a normalidade das classes.

Alguns classificadores baseados em aprendizado de máquina, como florestas aleatórias e redes neurais artificiais, apresentam estrutura mais robusta, sendo aplicáveis a situações de classificação que requerem dados heterogêneos, uma vez que derivam as premissas do modelo diretamente dos dados de treinamento, considerando que as classes não são modeladas por distribuições de probabilidade gaussianas. Desse modo, é possível reduzir significativamente a influência de dados não representativos das classes de interesse fornecendo resultados mais confiáveis.

As redes neurais artificiais (RNAs) são algoritmos multivariados e não paramétricos, inspirados na estrutura neural de organismos inteligentes em que seus elementos são neurônios organizados em camadas interconectadas. Sua modelagem é baseada em dados estatísticos não lineares e possui capacidade de processar dados complexos. A estrutura de uma RNA é composta, basicamente, por uma camada de entrada, onde os padrões são apresentados à rede, camadas intermediárias ou escondidas onde, por meio de conexões ponderadas, são realizados os processamentos e as características são extraídas, e a camada de saída, que apresenta o resultado final (BENEDIKTSSON et al., 1990).

O elemento computacional básico de uma RNA é um neurônio modelo (ou nó), o qual recebe informações, podendo ser de outros nós ou de dados de entrada, e as transforma em dados de saída (Figura 8). Em linhas gerais, os dados de entrada (X) são apresentados à rede e cada entrada é multiplicada por um peso (W) inicializado com um valor aleatório, que indica a sua influência na saída do nó, em seguida, é feita uma soma ponderada desses dados formando um conjunto de entrada. Esse resultado passa por uma função de ativação linear transmitindo o resultado à camada de saída. Se esse resultado exceder o limite da função de ativação, o neurônio será ativado e retornará um valor. Se a resposta de saída for considerada errada, o algoritmo de aprendizado ajusta os pesos para mapear os dados de entrada corretamente para as classes pré-definidas. Então, o próximo padrão de entrada é apresentado a rede e o processo é repetido para todos os padrões de entrada até que o aprendizado seja concluído (LARY, 2010).

Figura 8 - Estrutura de um neurônio artificial

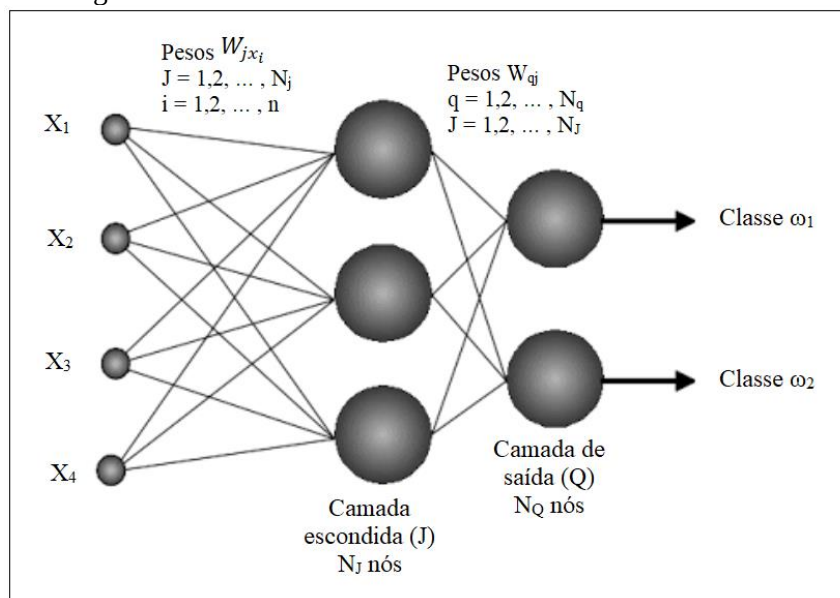


Fonte: Adaptado de Richards e Jia (2006)

A complexidade de uma RNA depende da quantidade de camadas escondidas e da quantidade de neurônios dessas camadas. A dimensionalidade dos dados de entrada define o número de neurônios da camada de entrada e o número de classes de interesse pré-definidas define o número de neurônios da camada de saída.

Na Figura 9 é apresentado um exemplo de modelo de RNA com duas camadas: uma camada intermediária J e uma camada de saída Q. Os atributos X1, X2, X3 e X4 são os conjuntos de dados de entrada; W_{jxi} e W_{qj} representam os pesos que são ajustados durante o treinamento; e por fim, ω_1 e ω_2 são as classes de saída da rede.

Figura 9 - Modelo de rede neural com uma camada escondida



Fonte: Adaptado de Gonzalez e Woods (2010)

O funcionamento de uma rede neural, conforme a apresentada na Figura 8, pode ser resumido nas seguintes etapas: inicialmente os atributos (X) são apresentados à camada de entrada (N). A entrada de cada neurônio da camada seguinte é resultado da

soma ponderada das saídas da camada anterior. Assim, a entrada (I_q) de cada neurônio na camada Q , é dada por:

$$I_q = \sum_{j=1}^{N_j} W_{qj} O_j \quad (2)$$

para $j=1, 2, \dots, N_j$, onde N_j é o número de neurônios da camada J e W_{qj} são os pesos que modificam as saídas O_j . As saídas da camada J são definidas através da expressão:

$$O_j = h_j(I_j) \quad (3)$$

para $j = 1, 2, \dots, N_j$.

Quando a função de ativação de uma rede neural é definida como sendo sigmóide, cada elemento de processamento da rede neural passa a ser definido pela seguinte equação:

$$h_q(I_q) = \frac{1}{1 + e^{-\left(\sum_{j=1}^{N_j} \omega_{qj} O_j + \theta_q\right) \theta_0}} \quad (4)$$

na qual $h_q(I_q)$ é a forma da função de ativação sigmoidal e Θ_q é um ganho adicional acrescido ao somatório dos coeficientes.

Um dos algoritmos mais utilizados no treinamento de uma rede neural com múltiplas camadas é o *backpropagation* (retropropagação). Essa técnica corrige os valores dos pesos pela diferença entre o valor obtido e o esperado no momento do treinamento e esse valor é propagado a partir da camada de saída até a camada de entrada, e então, os pesos das conexões das camadas internas vão sendo modificados (RICHARDS e JIA, 2006). Os passos gerais desse processo são apresentados a seguir e o seu funcionamento completo descrito em GONZALEZ e WOODS (2010).

O erro quadrático total (E_Q) entre as saídas desejadas (r_q) e as saídas obtidas (O_q) nos neurônios na camada de saída Q , é dado por:

$$E_q = \frac{1}{2} \sum_{q=1}^{N_Q} (r_q - O_q)^2 \quad (5)$$

onde N_Q é o número de neurônios na camada de saída Q.

Considerando que os pesos (W_{qj}) são ajustados na proporção da derivada parcial do erro, tem-se:

$$\Delta W_{qj} = -\alpha \frac{\partial E_Q}{\partial W_{qj}} \quad (6)$$

em que α é uma constante positiva, denominada taxa de aprendizado, e que determina a proporção de mudança nos pesos calculados para correção desses pesos.

Então, é utilizada a regra da cadeia para expressar a derivada parcial em termos de taxa de mudança de E_Q em relação à O_Q e a taxa de mudança de O_Q em relação a I_Q :

$$\Delta W_{qj} = -\alpha \frac{\partial E_Q}{\partial I_Q} O_j \quad (7)$$

ou

$$\Delta W_{qj} = -\alpha \delta_q O_j \quad (8)$$

em que

$$\delta_q = -\frac{\partial E_q}{\partial I_q} \quad (9)$$

Para a camada de saída Q, o termo de erro δ_q passa a incorporar a função de ativação da Equação 4:

$$\delta_q = (r_q - O_q) h'_q(I_q) \quad (10)$$

sendo h'_q a função de ativação, definida para cada elemento de processamento da rede.

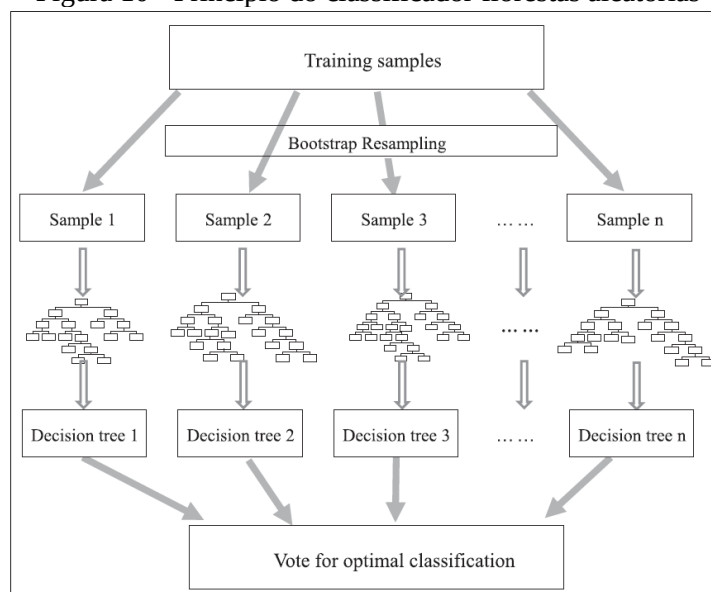
Considerando a propagação do erro de volta para o ajuste dos pesos dos nós da camada escondida J, a partir da camada de saída Q, tem-se a seguinte expressão para o valor esperado de δ_j :

$$\delta_j = h'_j(I_j) \sum_{q=1}^{N_Q} \delta_q W_{jq} \quad (11)$$

Uma sessão de treinamento bem-sucedida, permite que o erro da rede diminua com o número de iterações, convergindo para um conjunto estável de pesos. Assim, o procedimento para determinar se um padrão foi classificado corretamente durante o treinamento baseia-se em determinar se a resposta do nó na camada de saída, é alta, enquanto todos os outros nós apresentam resposta baixa (GONZALEZ e WOODS, 2010).

Outro classificador baseado em aprendizado de máquina que tem recebido crescente atenção da comunidade de sensoriamento remoto devido aos excelentes resultados de classificação obtidos e à facilidade de processamento, é o florestas aleatórias (FA), proposto por Breiman (2001). FA é um algoritmo de aprendizado supervisionado que faz uso de um grande número de árvores de decisão para superar as dificuldades de uma única árvore. Consiste numa coleção de classificadores estruturados em árvore, em que cada classificador é gerado usando um vetor aleatório amostrado independentemente do vetor de entrada. Assim, cada árvore lança um voto unitário e o voto “majoritário” de todas as árvores é usado para atribuir uma classe final para o vetor de entrada (BREIMAN, 2001). A Figura 10, mostra basicamente, a estrutura adotada pelo algoritmo FA.

Figura 10 - Princípio do classificador florestas aleatórias



Fonte: Zhang et al. (2017)

O classificador FA não necessita de informações sobre a distribuição dos dados e é mais robusto em relação ao ruído. Além disso, uma qualidade do classificador por florestas aleatórias, é o fato de fornecer estimativas internas que informam a importância relativa de cada atributo para uma dada situação de classificação. Essas estimativas auxiliam no processo de escolha dos atributos que mais contribuem para a predição, evitando assim, um *overfitting*.

Alguns parâmetros que devem ser definidos pelo usuário, de acordo com o classificador, podem afetar a precisão da classificação. Embora a literatura sugira valores padrões, testes empíricos são essenciais para garantir que a melhor classificação possível seja realizada. Maxwell et al. (2018) mostra que para uma RNA treinada com *backpropagation*, os diversos parâmetros a serem definidos se tornam um desafio. Esses parâmetros são: função de ativação, número de camadas ocultas, número de nós em cada camada, taxa de aprendizado, fator de momento, inicialização de peso, número de iterações e parâmetros de *pruning*. Por outro lado, o FA é considerado de fácil otimização em comparação com outros algoritmos de aprendizado de máquina, como a RNAs, pois requer a definição de apenas dois parâmetros pelo usuário: o número de árvores e o número de variáveis amostradas aleatoriamente como candidatas em cada divisão.

O tamanho e qualidade das amostras de treinamento também são questões-chave e devem ser consideradas. O problema dos dados de treinamento em desequilíbrio (desbalanceados) é comum em aplicações utilizando técnicas sensoramento remoto, em que os tipos de coberturas da terra mais raras são sub-representados em relação a classes

mais abundantes, devido a restrições de tempo e recursos para a coleta de amostras de treinamento representativas que sejam suficientes. Estudos mostraram que conjuntos de dados balanceados melhoram a classificação geral em comparação com dados desequilibrados (ESTABROOKS et al., 2004; WEISS e PROVOST, 2003).

Existem diversas técnicas alternativas que podem ser empregadas para equilibrar a quantidade de dados de treinamento por classe. Isso passa pela subamostragem das classes majoritárias (FREEMAN et al., 2012) e maior ponderação dos dados de treinamento das classes menos representativas (CHEN et al., 2004). Há também técnicas que contornam o problema da baixa amostragem da classe minoritária por meio da replicação de suas amostras para corresponder à quantidade de amostras de treinamento da classe majoritária (LING e LI, 1998).

Estudos indicam que o FA sofre grande influência com a utilização de dados de treinamento desbalanceados. Um estudo desenvolvido por Stumpf et al. (2011), mostrou que, ao utilizar amostras desequilibradas de áreas com deslizamento de terra, onde o número de amostras de locais sem deslizamento era muito maior que o de locais com deslizamento, o classificador FA subestimou a ocorrência de deslizamentos, prejudicando o seu desempenho geral.

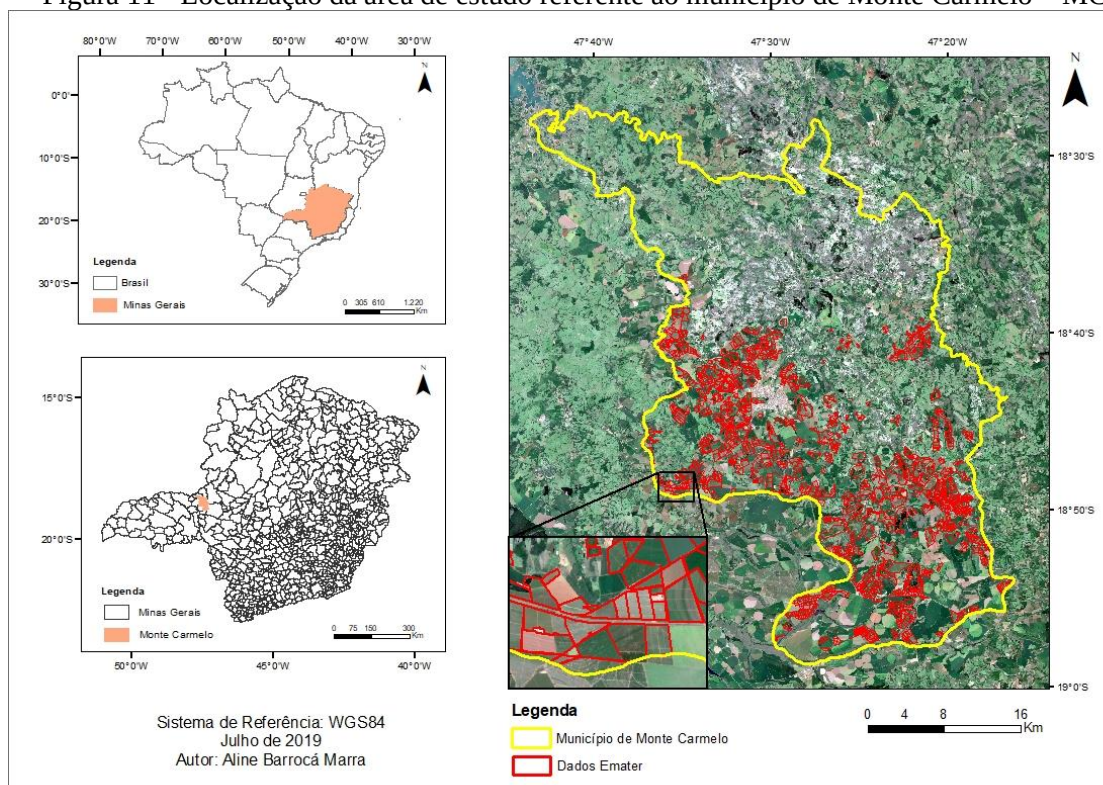
5. MATERIAIS E MÉTODOS

5.1 Materiais

5.1.1 Área de estudo

A área de estudo está localizada na região sul do estado de Minas Gerais, mais precisamente no município de Monte Carmelo (Figura 11). O município de Monte Carmelo apresenta uma forte extensão agrícola, onde a produção cafeeira é predominante, representando aproximadamente 30% da produção nacional (MARTINS, 2017). O município está inserido em região do Cerrado Mineiro, a qual possui solos arenosos que, independente do clima, caracterizam-se, em geral, pela sua acidez, porosidade, profundidade, baixa retenção de água e são pobres em alguns nutrientes.

Figura 11 - Localização da área de estudo referente ao município de Monte Carmelo – MG



A Empresa de Assistência Técnica e Extensão Rural do Estado de Minas Gerais (Emater - MG) é uma empresa pública brasileira vinculada à Secretaria de Estado de Agricultura, Pecuária e Abastecimento, do governo do Estado de Minas Gerais e é

responsável pelo atendimento a aproximadamente 400 mil agricultores mineiros. A empresa possui um geoportal (geoportal do café) que oferece informações precisas sobre o setor. Entre as informações disponibilizadas, são oferecidos dados vetoriais de áreas mapeadas de plantações de café de todo o estado de Minas Gerais (formato .kml). Trata-se de polígonos indicando os talhões de café assistidos pela empresa (polígonos em vermelho da Figura 11). Os dados podem ser encontrados na página <http://geoportaldocafe.emater.mg.gov.br/ferramenta>.

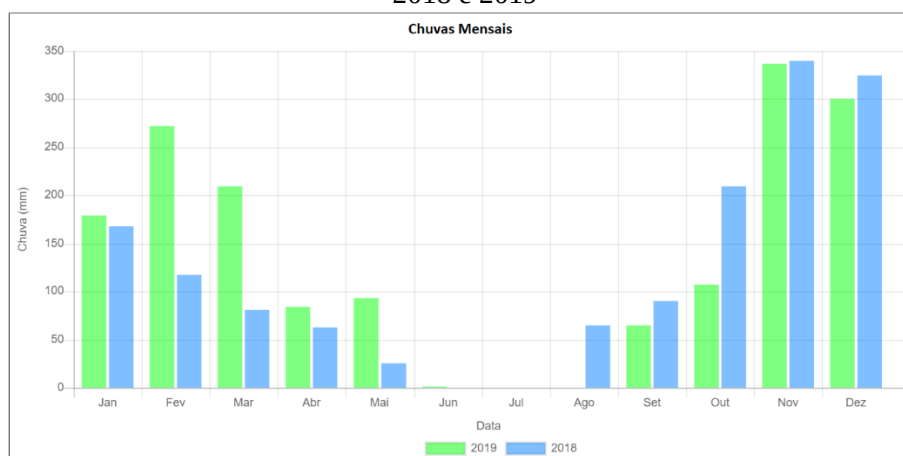
O município de Monte Carmelo possui altitude média de 825 metros e uma área total de 1.344 km². O cultivo de café no município compreende uma área de 18.549 hectares, na qual predomina a produção de café arábica irrigado, conforme indicado na Tabela 2. A Figura 12 e a Figura 13 mostram, respectivamente, as precipitações e temperaturas mensais médias do município para os anos de 2018 e 2019.

Tabela 2 - Variedades de café produzidos no município de Monte Carmelo - MG

Variedade de Café	Ano	Área (hec.)	Produção (sacas/ano)	Produtividade (sacas/hec./ano)
Café Arábica Sequeiro	2016	12.690,00	507.600,00	40,00
	2017	190,00	6.080,00	32,00
	2018	190,00	6.080,00	32,00
Café Arábica Irrigado	2017	12.500,00	550.000,00	44,00
	2018	12.500,00	550.000,00	44,00

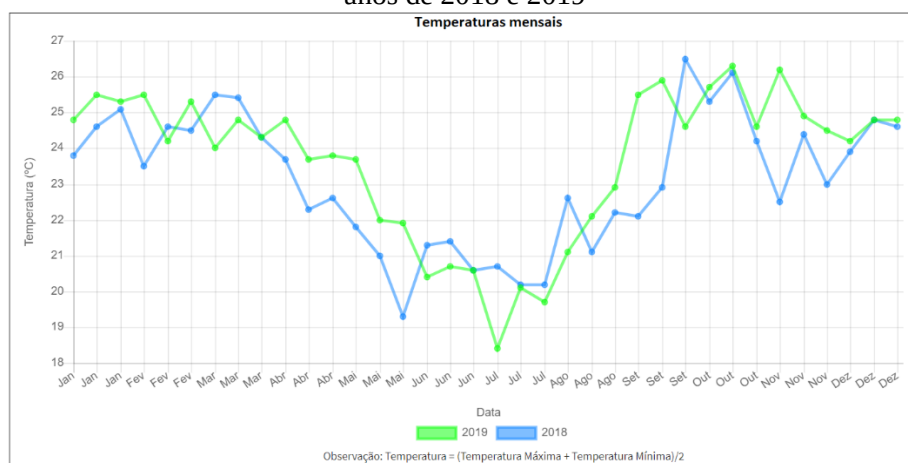
Fonte: EMATER – MG, 2019

Figura 12 – Chuva mensal média do município de Monte Carmelo – MG, referente aos anos de 2018 e 2019



Fonte: Sismet Cooxupé, 2020

Figura 13 – Temperatura mensal média do município de Monte Carmelo – MG, referente aos anos de 2018 e 2019



Fonte: Sismet Cooxupé, 2020

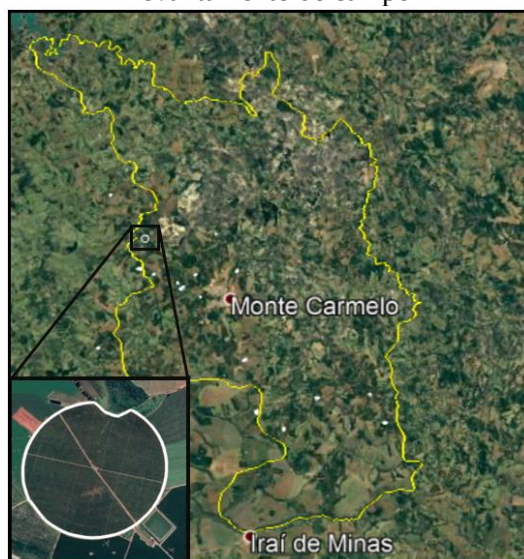
Como pode ser visto na Figura 12, nos meses de verão as chuvas são mais acentuadas, com maior precipitação nos meses de novembro e dezembro, apresentando médias entre os anos de 2018 e 2019, de 338,3 mm (novembro) e 312,5 mm (dezembro). As menores precipitações aparecem em junho (média dos anos de 2018 e 2019 de 0,5 mm) e em julho (média dos anos de 2018 e 2019 de 0,0 mm). Em relação à temperatura do município (Figura 13), o mês mais frio é julho (média dos anos de 2018 e 2019 de 19,5 °C) e o mais quente outubro (média dos anos de 2018 e 2019 de 25,4 °C).

O presente estudo será direcionado às áreas que compreendem o mapeamento de café realizado pela Emater-MG para a investigação do problema abordado. Além disso, em parceria com a Universidade Federal de Uberlândia (UFU), foram disponibilizadas localizações de talhões de café no cerrado mineiro com a presença constatada de nematoides, inseridas nas áreas mapeadas pela Emater-MG.

5.1.2 Dados de referência

Inicialmente, foi disponibilizada pela Universidade Federal de Uberlândia a localização de um primeiro conjunto talhões de café infectado por nematoides, referente a um primeiro levantamento de campo realizado (Figura 14). Para a realização dos primeiros experimentos, as amostras de café infectado foram coletadas com base nesses dados disponibilizados.

Figura 14 – Amostras de café infectado disponibilizadas pela UFU referente ao primeiro levantamento de campo



Fonte: Google Earth

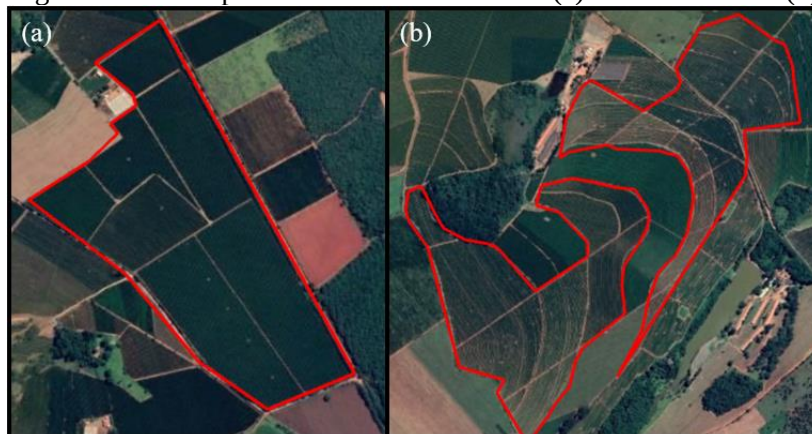
Posteriormente, foi realizado um novo levantamento de campo e produzido um segundo conjunto de dados de referência (Figura 15) contendo amostras de café infectado e café sadio. Nesse caso, trata-se somente de talhões de café irrigado. A Figura 16 mostra alguns exemplos de talhões de café infectado e sadio disponibilizados.

Figura 15 – Localização das amostras de café infectado e sadio do segundo conjunto de dados disponibilizados pela UFU



Fonte: Google Earth

Figura 16 – Exemplo de talhões de café sadio (a) e café infectado (b)



Fonte: *Google Earth*

Para os demais experimentos, foram utilizadas somente as amostras referentes ao segundo levantamento de campo realizado pela UFU. As amostras foram coletadas com base nesses dados disponibilizados, assim como, abrangendo as áreas mapeadas e assistidas pela EMATER – MG.

Os dados disponibilizados nos dois levantamentos de campo, foram utilizados para verificar o potencial da região espectral do SWIR (infravermelho de ondas curtas), na investigação de áreas infectadas por nematoides.

5.1.3 Imagens multiespectrais MSI/Sentinel-2

Copernicus Sentinel-2 é uma missão de imageamento multiespectral que compreende uma constelação de dois satélites em órbita polar sincronizada com o Sol, divididos em 180° entre si. Destina-se a monitorar a variabilidade nas condições da superfície da terra com uma faixa orbital de 290 km e possui revisita de dez dias no Equador com um satélite e cinco dias com dois satélites. Os limites de cobertura são entre as latitudes 56° sul e 84° norte. Os satélites Sentinel-2A foram lançados em 23 de junho de 2015 e os Sentinel-2B em 07 de março de 2017 e estão equipados com o instrumento sensor MSI (*MultiSpectral Instrument*), que oferece imagens ópticas de alta resolução (ESA, 2019).

O Sentinel-2 possui resolução radiométrica de 12 bits e dispõe de 13 bandas espectrais em três diferentes resoluções espaciais: quatro bandas a 10 m, seis bandas a 20 m e três bandas a 60 m, como pode ser visto na Tabela 3:

Tabela 3 - Bandas espectrais do Sentinel-2

Resolução espacial (m)	Número da banda	Nome da banda	S2A	S2B
			Comprimento de onda central (nm)	
10	2	Azul	492,4	492,1
	3	Verde	559,8	559
	4	Vermelho	664,6	664,9
	8	Infravermelho Próximo	832,8	832,9
20	5	Red Edge 1	704,1	703,8
	6	Red Edge 2	740,5	739,1
	7	Red Edge 3	782,8	779,7
	8A	Red Edge 4	864,7	864
	11	Infravermelho de Ondas Curtas 1	1613,7	1610,4
	12	Infravermelho de Ondas Curtas 2	2202,4	2185,7
60	1	Aerossol	442,7	442,2
	9	Water Vapor	945,1	943,2
	10	Cirrus	1373,5	1376,9

Fonte: ESA, 2019

As imagens do Sentinel-2 estão disponíveis para usuários em diferentes níveis de correção. O nível 1C refere-se a imagem com pixels expressos em reflectância de topo de atmosfera e o nível 2A em reflectância de superfície. O produto de nível 2A pode ser adquirido pelo usuário, tanto pelo *download* na página da Agência Espacial Europeia (ESA), quanto pelo processamento utilizando uma ferramenta criada pela agência chamada *sen2cor*, que utiliza como base o produto nível 1C. Os produtos Sentinel-2 de níveis 1C e 2A são ortoimagens de 100 km² na projeção/Datum UTM/WGS84.

Devido à grande quantidade de dados gratuitos do sistema Sentinel-2 e de suas características, há uma grande oportunidade de monitorar colheitas em tempo quase real. A resolução espacial da missão Sentinel-2, permite trabalhar com diversas bandas espectrais e também gerar um grande número de imagens derivadas para investigar o problema abordado.

5.1.4 Ferramentas Computacionais

As ferramentas utilizadas para o desenvolvimento da pesquisa são os *softwares* ENVI, QGIS, ArcGIS, SNAP e o sistema de aprendizado de máquina WEKA:

ENVI 5.1:

ENVI (*Environment for Visualizing Images*), desenvolvido pela *Harris Geospatial Solutions*, é um sistema de processamento de imagens que permite a visualização, exploração, análise e apresentação de dados na área de sensoriamento remoto combinando ferramentas de alto desempenho e apresenta facilidade de uso. Sua arquitetura permite trabalhar com imagens fornecidas por sensores de última geração como IKONOS, QuickBird, WorldView, Sentinel, entre outros.

Neste estudo, ele é utilizado para a coleta de amostras de treinamento das classes de interesse pois permite a exportação de amostras em formato ASCII, o qual fornece a caracterização das amostras dos planos de entrada.

QGIS 3.6.3:

Gary Sherman começou o desenvolvimento do *Quantum GIS* no início de 2002, e logo tornou-se um projeto incubado no *Open Source Geospatial Foundation* em 2007, onde sua primeira versão foi lançada em janeiro de 2009. O QGIS é um sistema de informação geográfica livre e de código aberto que permite criar, editar, visualizar e analisar informação geoespacial em diversas plataformas. Nesse estudo, ele é usado para o cálculo dos índices espectrais. Está disponível para *download* na página https://www.qgis.org/pt_BR/site/forusers/download.html.

ArcGIS 10.6.1:

Devido a necessidade de analisar um número cada vez maior de projetos, a Esri desenvolveu o ARC/INFO, o primeiro produto GIS comercial, lançado em 1981. O ArcGIS é um sistema de informação geográfica que oferece um conjunto de funcionalidades baseadas em localização para diversas análises. Foi utilizado na geração de mapas temáticos resultantes dos processamentos realizados no *software* WEKA.

SNAP 6.0.0:

Desenvolvido pela agência Espacial Europeia (ESA), o *Sentinel Application Platform* (SNAP), reúne todas as caixas de ferramentas do Sentinel para oferecer uma plataforma mais complexa. É um aplicativo de processamento de imagens e possui algumas funções básicas como abrir um produto, explorar os componentes do produto como bandas e máscaras, explorar informações dos pixels, gerar composições coloridas,

realizar operações entre bandas, calcular índices espectrais, realizar a classificação de imagens, entre diversas outras funções.

O programa é livre e trabalha com outros sensores além do MSI/Sentinel, como Landsat, RapidEye, Spot, WorldView e etc. Está disponível para *download* na página <http://step.esa.int/main/download/> e tutoriais para diversas aplicações do programa podem ser encontrados na página <https://step.esa.int/main/doc/tutorials/snap-tutorials/>. O sistema foi utilizado para alguns processamentos nas imagens Sentinel-2 adquiridas, como reamostragem dos pixels, correção atmosférica e recorte da imagem original. Além disso, ele também foi utilizado para gerar imagens de textura.

Weka 3.8.3:

O projeto de software WEKA (*Waikato Environment for Knowledge Analysis*) começou a ser desenvolvido em 1993, utilizando a linguagem Java, na Universidade de Waikato, financiado pelo governo da Nova Zelândia e está dentro das especificações da GPL (*General Public License*). É um software livre de código aberto e foi desenvolvido devido à necessidade de um banco de trabalho unificado, que permitisse aos pesquisadores acesso fácil a técnicas de última geração no aprendizado de máquina. Foi consolidado como a ferramenta de mineração de dados mais utilizada por estudantes e professores de universidades (HALL et al., 2009).

O sistema WEKA tem o objetivo de agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo de aprendizado de máquina. Ele também permite a manipulação das configurações dos algoritmos baseados em aprendizado de máquina para a classificação dos dados. Nesse estudo, esse sistema é utilizado para a análise da importância dos atributos e para a classificação das imagens. O sistema pode ser obtido pela página <https://www.cs.waikato.ac.nz/ml/WEKA/>.

5.2 Metodologia

Para avaliar a capacidade de dados multiespectrais Sentinel-2 combinados com algoritmos baseados em aprendizado de máquina na detecção de nematoides em produções de café, algumas etapas foram seguidas.

A primeira etapa consistiu numa revisão teórica detalhada sobre o problema abordado. Após a revisão, foram adquiridas imagens Sentinel-2. Inicialmente, foi adquirida uma imagem de uma única data, referente ao mês de julho de 2018, em reflectância de topo de atmosfera (nível 1C), convertida posteriormente em reflectância de superfície (nível 2A). Em seguida, foram adquiridas outras duas imagens, do nível 2A, referentes aos meses de janeiro e abril de 2019, com o objetivo de incluir as principais fases fenológicas do café em um segundo experimento desenvolvido. No passo seguinte, foram analisadas as características dos talhões de café infectados disponibilizados pela UFU, assim como das áreas não infectadas, referentes aos dados disponibilizados a partir dos dois levantamentos de campo realizados.

O primeiro experimento foi realizado utilizando a imagem Sentinel-2 do mês de julho de 2018, a partir da qual foram calculados e avaliados os índices de vegetação, que discriminem os talhões infectados por nematoides, por meio da aplicação de algumas métricas tradicionais. Para isso, foram definidas amostras de referência na imagem, identificadas com base nos dados do primeiro levantamento de campo disponibilizados pela UFU.

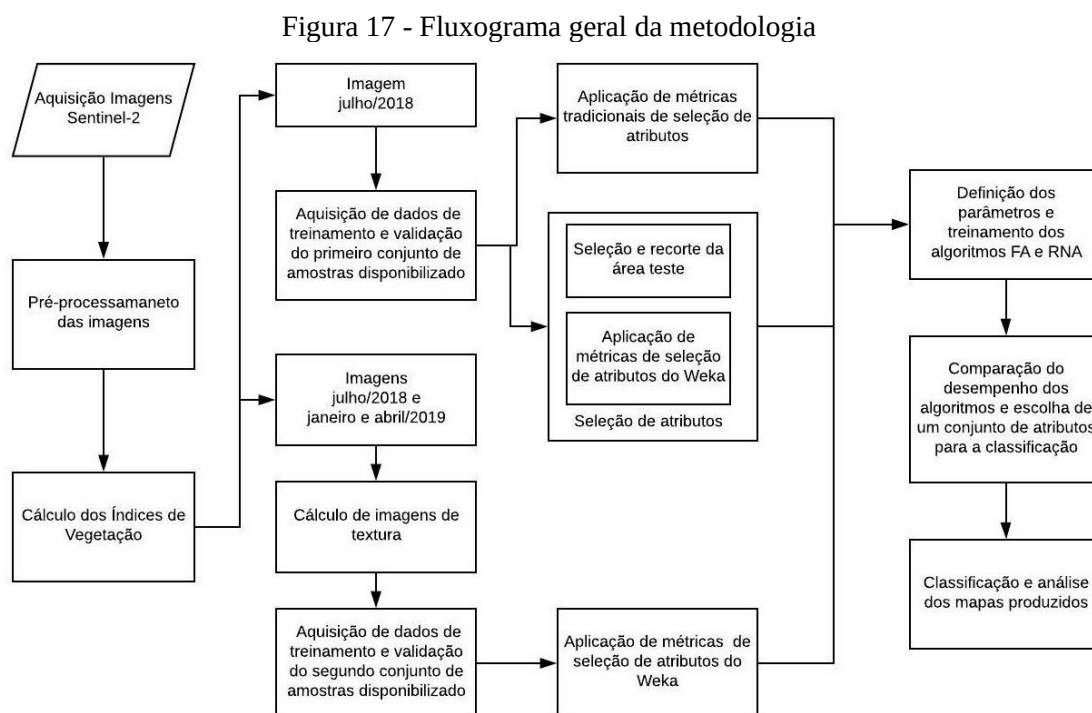
Ainda neste primeiro experimento, uma área teste foi selecionada e recortada da imagem original de 2018 e coletadas novas amostras de áreas de café infectado, café sadio e solo exposto somente de dentro dessa área teste na imagem, também com base no primeiro conjunto de amostras disponibilizado. Posteriormente foram aplicadas algumas métricas de seleção de atributos presentes no *software* WEKA, com o objetivo de gerar indicadores da importância de cada atributo.

Para o desenvolvimento do segundo experimento foram coletadas novas amostras que constituíram outro conjunto de dados disponibilizados pela UFU e utilizadas as imagens das três datas (julho/2018; janeiro/2019 e abril/2019). Além disso, foram aplicadas as métricas de seleção de atributos do WEKA definidas para o primeiro experimento. Nesse segundo caso, além dos índices de vegetação, também foram incluídas imagens de textura aos dados de entrada.

Os dois experimentos desenvolvidos auxiliaram na seleção de alguns conjuntos de atributos que apresentaram maior eficiência da discriminação entre as classes de café infectado e sadio. Esses atributos foram utilizados como dados de entrada no processo de treinamento dos algoritmos baseados em aprendizado de máquina RNA e FA (disponíveis no WEKA). A classificação foi desenvolvida apenas para os talhões de café, definidos segundo uma máscara contendo as áreas de café assistidas pela Emater-MG. A utilização

da máscara da cultura garante que o processo seja realizado apenas nas áreas cultivadas com café. Os dados de treinamento de talhões de café infectado e sadio utilizados para a classificação foram extraídos do segundo conjunto de dados de campo disponibilizados pela UFU.

O processo de classificação constou da aplicação dos algoritmos RNAs e FA para a discriminação de áreas de café infectadas por nematoides e, posteriormente, foi feita uma análise comparativa do desempenho desses classificadores. Em seguida, foram gerados mapas da distribuição espacial das áreas infectadas e uma estimativa da qualidade desses mapas por meio da representação espacial da incerteza da classificação. A estrutura geral do projeto é apresentada na Figura 17:



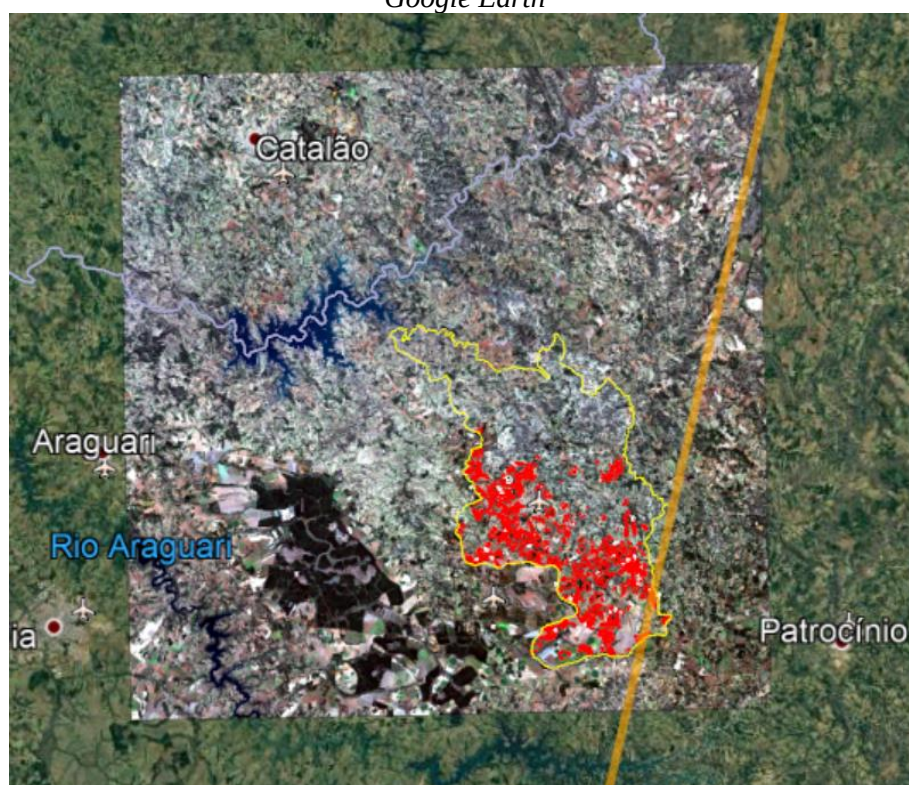
5.2.1 Aquisição dos dados multiespectrais MSI/Sentinel-2

O *Copernicus Open Access Hub* fornece acesso completo, gratuito e aberto aos produtos Sentinel para os usuários por meio de uma ferramenta chamada *Open Hub*. A ferramenta pode ser encontrada na página <https://scihub.copernicus.eu>. Dados Sentinel-1, Sentinel-2 e Sentinel-3 podem ser adquiridos pela ferramenta.

Para o *download* das imagens Sentinel-2, devem ser especificados o satélite (S2A ou S2B), o tipo do produto (níveis 1C ou 2A), a porcentagem de cobertura de nuvem admissível, o número relativo de órbita, a data e a área desejada.

Inicialmente, para a realização do primeiro experimento, foi adquirida uma imagem referente a julho de 2018 em nível 1C (reflectância de topo de atmosfera), devido a não disponibilidade da imagem em nível 2A (reflectância de superfície). A imagem foi tomada pelo satélite 2A no dia 19 de julho de 2018 com sentido descendente e número relativo da órbita 38, como pode ser visto na Figura 18.

Figura 18 – Inserção da imagem Sentinel-2A da área de estudos, sobreposta à visualização do *Google Earth*



Fonte: *Google Earth*

A Figura 18 mostra a órbita relativa 38 limitada à direita pela linha em cor laranja, os dados vetoriais de culturas de café disponibilizados pela Emater-MG em vermelho, os limites do município de Monte Carmelo-MG em amarelo e a imagem Sentinel-2A adquirida em composição colorida verdadeira (RGB-432).

Após o segundo levantamento de campo realizado pela UFU, foram adquiridas duas outras datas de tomada de imagens, abrangendo outras fases fenológicas do café para a realização do segundo experimento, permitindo uma melhor investigação de áreas de café infectado na região de estudo. As imagens adquiridas foram tomadas nos dias 30 de

janeiro e 20 de abril de 2019, com sentido descendente e número relativo da órbita 38. Ambas as imagens foram tomadas pelo satélite S2B em reflectância de superfície (nível 2A).

5.2.2 Pré-processamento das imagens

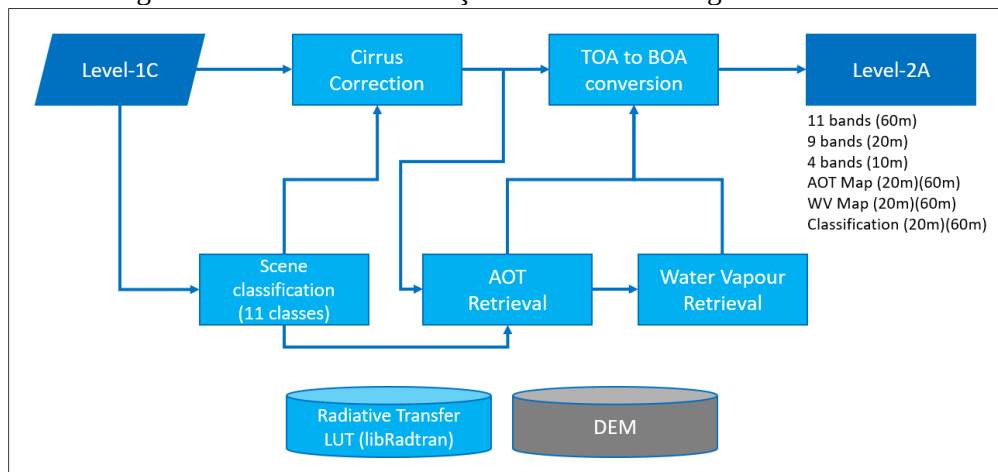
O processamento digital das imagens envolveu correção atmosférica, reamostragem e recorte das imagens utilizando o aplicativo computacional SNAP.

Correção atmosférica:

Como dito anteriormente, a imagem Sentinel-2A de julho de 2018 foi adquirida no nível 1C e foi convertida em nível 2A por meio da ferramenta *Sen2Cor*. Entre outras opções, o nível 2A pode ser gerado pelo usuário a partir do produto de nível 1C usando a caixa de ferramentas do Sentinel-2 presente no *software* SNAP ou pela versão independente do processador oferecida gratuitamente na página <https://step.esa.int/main/third-party-plugins-2/sen2cor/>.

O processamento de nível 2A inclui uma classificação de cena e uma correção atmosférica aplicada aos produtos de ortoimagem de nível 1C em topo de atmosfera (*Top Of Atmosphere* - TOA). A saída principal de nível 2A é um produto de reflectância corrigido de base de atmosfera (*Bottom Of Atmosphere* - BOA) ou de superfície. Para o processamento, foi utilizada o algoritmo *Sen2Cor* disponível no SNAP e foram adotados o modelo atmosférico *mid-latitude winter* e o modelo de aerossol rural. Os demais parâmetros foram utilizados como *default*.

Os produtos de saída nível 2A são um mapa de classificação de cena (SCL) juntamente com indicadores de qualidade para probabilidades de nuvens e neve, mapas *True Color Image* (TCI), *Aerosol Optical Thickness* (AOT) e *Water Vapor* (WV) e imagens de reflectância de superfície que são fornecidas em diferentes resoluções espaciais (10, 20 e 60m). A estrutura básica do processador *Sen2Cor* pode ser vista na Figura 19:

Figura 19 - Processo de correção atmosférica do algoritmo *Sen2Cor*

Fonte: Louis et al. (2016)

Main-Knorn et al. (2017) desenvolveram um estudo sobre validações envolvendo a correção atmosférica realizada pelo algoritmo *Sen2Cor* e verificaram que os resultados foram promissores apresentando valores de RMSE entre a reflectância de superfície obtida utilizando o aplicativo *Sen2Cor* e medições *in situ* variando de 0,002 a 0,005.

Essa etapa não foi necessária para as imagens de janeiro e abril de 2019, pois elas já foram adquiridas em nível 2A, ou seja, em reflectância de base de atmosfera (*Bottom Of Atmosphere* - BOA) ou de superfície.

Reamostragem das imagens:

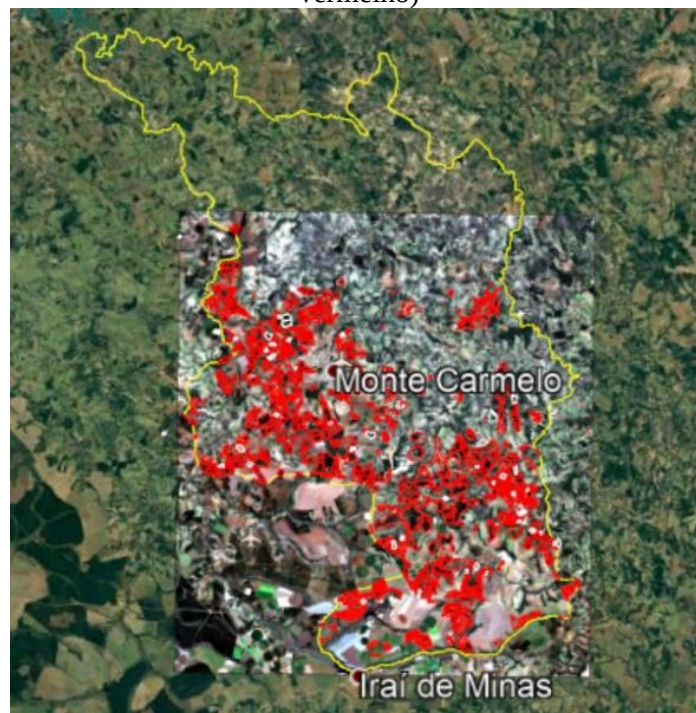
Devido às diferentes resoluções espaciais das imagens Sentinel-2, foi necessário realizar a reamostragem para padronizar o tamanho dos pixels nos intervalos espectrais distintos. Para esse processamento foram utilizadas as imagens resultantes do processo de correção atmosférica (imagens nível 2A) da imagem de julho/2018 e as imagens originais de janeiro/2019 e abril/2019. As imagens foram reamostradas para 10 m de resolução espacial utilizando a ferramenta *resampling* presente no SNAP. Foi utilizada a banda B02 (azul) como referência (banda originalmente com resolução de 10 m) e o método de interpolação *Nearest*, baseado no vizinho mais próximo (*nearest neighbour*).

Recorte das imagens:

Apesar da necessidade de aplicar posteriormente uma máscara da cultura para isolar os talhões de café, a imagem foi inicialmente recortada com a ferramenta *Subset* do SNAP. O recorte foi feito de forma que abrangesse somente a área de café mapeada pela

Emater-MG, de modo a reduzir o tamanho da cena e otimizar a operação dos algoritmos necessários para alguns processamentos, como por exemplo, a definição da matriz de co-ocorrência. A área recortada da imagem é mostrada na Figura 20.

Figura 20 – Recorte da imagem abrangendo os talhões de café em estudo (delimitados em vermelho)



Fonte: Google Earth

5.2.3 Geração dos índices de vegetação e imagens de textura

Os índices de vegetação foram usados nos dois experimentos. A geração dos índices foi realizada no *software* QGIS 3.6.3 de acordo com a formulação apresentada na Tabela 1. Os índices utilizados foram: NDVI, NDWI, CSI, NBR2, MSI, MCARI*, TCARI* e NRI. Esses índices foram escolhidos pois incorporam uma banda na região do infravermelho de ondas curtas, sensível ao teor de umidade da planta. Como exposto anteriormente, quando infectada por nematoides, a raiz do cafeeiro pode perder a eficiência na absorção e translocação de água e nutrientes para a parte aérea da planta, de modo que foram utilizados os índices de vegetação cuja formulação incluía essa região do espectro.

As imagens de textura foram calculadas somente para os dados do segundo experimento. Elas foram obtidas por meio da ferramenta *Grey Level Co-occurrence Matrix* (GLCM) presente no SNAP. Os parâmetros adotados no cálculo da matriz de co-

ocorrência foram: tamanho da janela igual a 9 x 9 pixels; distância 1; nível de quantização 16 bits; e ângulos de 0°, 45°, 90° e 135° (opção *ALL*). Uma distância pequena e busca em todos os ângulos foram usados com o intuito de realizar uma boa captura da condição investigada.

Foram calculadas imagens de textura para as bandas do SWIR 1 e NIR das três datas: janeiro e abril/2019 e julho/2018. A banda do NIR foi escolhida por ser uma banda de resolução espacial original de 10m, para qual as bandas do SWIR e RE foram reamostradas.

5.2.4 Seleção de atributos

Independente do método de classificação supervisionada adotado, alguns procedimentos são essenciais para um adequado desempenho do algoritmo de classificação, entre eles, a escolha do melhor conjunto de bandas espectrais e outros atributos de natureza espacial, considerando o alvo de interesse (NOVO, 2010). Atributos que não possuem uma contribuição significativa podem deteriorar os resultados obtidos por esquemas baseados em aprendizado de máquina. Reduzir a dimensionalidade dos dados, excluindo atributos inadequados, melhora o desempenho dos algoritmos de aprendizado (WITTEN e FRANK 2005). Na busca de atributos que melhor discriminem a presença de nematoides em cafeeiros, foram realizados dois experimentos.

5.2.4.1 Primeiro experimento

No primeiro experimento, foi utilizada a imagem tomada no mês de julho de 2018 por ser considerado um período cujas condições são pouco favoráveis à detecção de talhões infectados por nematoides, já que se trata das fases de repouso e autopoda dentro do ciclo fenológico do café. Nesse período também ocorrem baixas temperaturas e precipitações, sendo possível observar, nessas condições, uma desfolha mesmo em áreas sadias, dificultando a discriminação de áreas de café infectado.

Por oferecer condições desfavoráveis à detecção de café infectado, a imagem de julho foi utilizada no primeiro experimento com o objetivo de verificar sua capacidade na detecção do problema, possibilitando, posteriormente, uma comparação dos resultados utilizando outras duas imagens tomadas em datas que compreendem outras fases fenológicas do cafeeiro, e conseqüentemente, outras condições.

Inicialmente, utilizando o *software* ENVI 5.1, foram coletadas amostras de café infectado e sadio com base no primeiro conjunto de dados disponibilizados pela UFU. Apesar das amostras terem sido coletadas na forma de polígonos, os dados foram exportados em formato ASCII (cuja configuração é exemplificada na Figura 21), o qual fornece os valores de cada pixel para cada banda (ou índice).

Figura 21 - Exemplo de parte dos dados de treinamento em arquivo ASCII

ID	X	Y	Map X	Map Y	Lat	Lon	B1	B2	B3
1	3525	2969	860920.00	7935180.00	-18.643419	-47.579679	0.0323	0.2905	0.3060
2	3526	2969	860930.00	7935180.00	-18.643417	-47.579584	0.0303	0.2905	0.2988
3	3524	2969	860910.00	7935180.00	-18.643421	-47.579774	0.0320	0.2967	0.3205
4	3519	2970	860860.00	7935170.00	-18.643520	-47.580245	0.0348	0.2885	0.3084
5	3520	2970	860870.00	7935170.00	-18.643518	-47.580150	0.0346	0.2885	0.3094
6	3521	2970	860880.00	7935170.00	-18.643516	-47.580056	0.0346	0.2912	0.3076
7	3522	2970	860890.00	7935170.00	-18.643514	-47.579961	0.0368	0.2912	0.3130
8	3523	2970	860900.00	7935170.00	-18.643513	-47.579866	0.0383	0.2967	0.3132
9	3524	2970	860910.00	7935170.00	-18.643511	-47.579772	0.0389	0.2967	0.3121
10	3525	2970	860920.00	7935170.00	-18.643509	-47.579677	0.0417	0.2905	0.3084
11	3531	2970	860980.00	7935170.00	-18.643499	-47.579109	0.0351	0.2967	0.3156
12	3517	2971	860840.00	7935160.00	-18.643613	-47.580433	0.0346	0.2960	0.3138
13	3518	2971	860850.00	7935160.00	-18.643611	-47.580338	0.0352	0.2960	0.3118

A primeira coluna do arquivo ASCII se refere ao número identificador de cada pixel. Da segunda a sétima coluna, são apresentadas as coordenadas de cada pixel em colunas e linhas da imagem, em coordenadas do sistema UTM e em coordenadas geográficas, respectivamente. Nas demais colunas estão os atributos que descrevem os pixels da imagem.

Dentre as amostras de café infectado por nematoides disponibilizadas pela UFU, foram utilizadas somente aquelas incluídas na região mapeada pela Emater-MG. A discriminação entre as classes de café infectado e sadio foi feita por meio da média dos valores de reflectância de cada pixel amostrado nas duas classes. Para identificar os intervalos espectrais que melhor separam as duas ocorrências, foi elaborado um gráfico com os valores obtidos para as classes de café infectado e sadio com base nas bandas espectrais disponíveis no sensor MSI/Sentinel-2. Os valores de média para as duas classes, seus respectivos desvios padrão e a diferença absoluta entre as médias, também foram considerados.

Após a análise das bandas espectrais, foram avaliadas as médias dos valores obtidos nos índices de vegetação. Com base na análise das bandas individuais e considerando que várias deficiências nutricionais reduzem os níveis de clorofila nas plantas (Evans, 1989) e ainda que a região do *Red Edge* 3 (RE3, 782,8 nm) sofre grande influência da concentração de clorofila (Jensen, 2009), alguns índices de vegetação foram

calculados em sua formulação original e outros pela substituição da banda do infravermelho próximo (NIR, 832,8 nm) pela do RE3. Sobre esse aspecto, Schuster et al. (2012) e Kim et al. (2014) relatam que a inclusão da região espectral do *Red Edge* pode fornecer informações adicionais para a classificação da vegetação.

As médias obtidas para os índices de vegetação foram dispostas em dois gráficos separados: índices não normalizados e normalizados. Adotou-se essa estratégia para que os valores dos índices não normalizados não dificultassem a visualização da separabilidade das médias das classes dos índices normalizados, que resultam em valores entre -1 e 1.

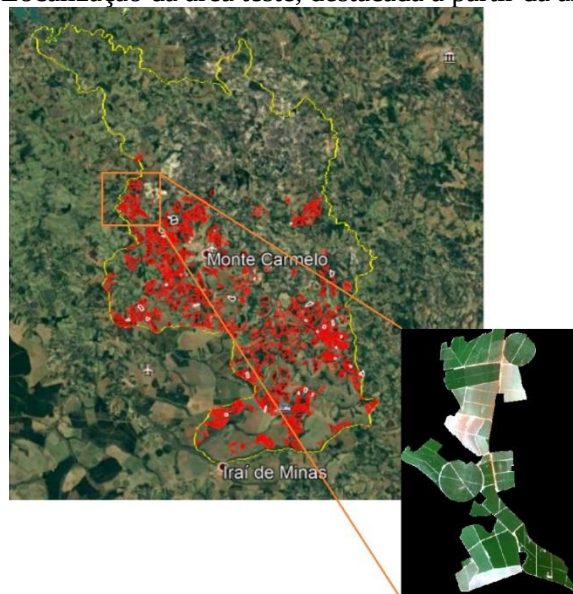
Além das médias dos atributos dos pixels, o índice M de separabilidade também foi utilizado para avaliar a discriminação entre as áreas de café infectado e sadio (KAUFMAN e REMER, 1994). Esse índice é utilizado para aferir a distinção entre classes com base na diferença absoluta da média dos valores das duas classes, dividido pela soma do desvio padrão (Equação 11):

$$M = \frac{|\mu_s - \mu_i|}{\sigma_s + \sigma_i} \quad (12)$$

em que μ_s (σ_s) e μ_i (σ_i) e são respectivamente a média (desvio padrão) dos valores do conjunto de pixels de café sadio e infectado. Essa média representa a sobreposição dos histogramas de frequência das classes analisadas, sendo que os valores de M maiores que 1 indicam alta separabilidade entre as classes, enquanto valores de M menores que 1, indicam baixa separabilidade.

Em seguida, foram aplicadas métricas para seleção dos atributos, selecionadas entre aquelas disponíveis no *software* WEKA que permitem avaliar a importância relativa de cada atributo. Para isso, somente uma pequena porção da área de estudo foi utilizada (em destaque na Figura 22) e novas amostras foram coletadas, com base no primeiro conjunto de dados disponibilizado.

Figura 22 - Localização da área teste, destacada a partir da área de estudos



A coleta das amostras de café infectado, café sadio e solo exposto foi feita por meio do aplicativo ENVI, posteriormente exportadas em ASCII e convertidas para o formato *Attribute - Relation File Format (ARFF)*, padrão do WEKA (exemplificado na Figura 23).

Figura 23 - Exemplo do arquivo ARFF contendo parte dos dados de treinamento

```
@relation Amostras

@attribute classe {cafe_infectado,cafe_sadio,solo_exposto}
@attribute B4 NUMERIC
@attribute B7 NUMERIC
@attribute B8 NUMERIC
@attribute B11 NUMERIC
@attribute B12 NUMERIC
@attribute NDVI_RE3 NUMERIC
@attribute NDWI_RE3 NUMERIC
@attribute NRI NUMERIC
@attribute NBR2 NUMERIC

@data
cafe_infectado,0.0323,0.2905,0.3060,0.1743,0.0940,0.8090,0.7999,0.2742,0.2500
cafe_infectado,0.0303,0.2905,0.2988,0.1743,0.0940,0.8159,0.8111,0.2632,0.2500
cafe_infectado,0.0320,0.2967,0.3205,0.1743,0.0949,0.8184,0.8053,0.2955,0.2599
cafe_infectado,0.0348,0.2885,0.3084,0.1783,0.0994,0.7972,0.7847,0.2673,0.2361
cafe_infectado,0.0346,0.2885,0.3094,0.1783,0.0994,0.7988,0.7858,0.2688,0.2361
cafe_infectado,0.0346,0.2912,0.3076,0.1768,0.0961,0.7978,0.7876,0.2700,0.2444
```

O arquivo ARFF é definido pelo cabeçalho e seção de dados. O cabeçalho contém a denominação das classes, seguido da lista de atributos e respectivos tipos, (numéricos, nominais e datas). A seção dos dados contém o nome de cada classe seguido dos valores de cada atributo daquela classe.

O arquivo ARFF foi utilizado como entrada para o cálculo das métricas de seleção de atributos no ambiente WEKA. O objetivo dessas métricas é pesquisar dentro do espaço de atributos disponível, o subconjunto com maior probabilidade de prever a classe para cada pixel. As métricas utilizadas foram: o *Classifier Attribute Evaluation*, o *Gain Ratio Attribute Evaluation* e o *Correlation Attribute Evaluation*.

O *Classifier Attribute Evaluation* (ClassAE) avalia o valor de um atributo usando um classificador especificado pelo usuário. Para esse estudo, foi selecionado o classificador Florestas Aleatórias, que fornece um critério de importância variável para cada atributo, calculando a redução média na acurácia da classificação, a partir de dados *out of bag* (OOB) da amostragem realizada por *bootstrap* (Verikas et al., 2011). O procedimento *bootstrap* é uma técnica de reamostragem bastante utilizada em diferentes aplicações estatísticas que permite a obtenção de um subconjunto de dados, por reamostragem do conjunto de dados original (Efron e Tibishirani, 1993).

Quando o treinamento do algoritmo Florestas Aleatórias é definido por amostragem com substituição, cerca de um terço dos dados é deixado de fora da amostra de treinamento. Esses dados OOB são usados para obter uma estimativa imparcial do erro de classificação à medida que as árvores são adicionadas à floresta.

O *Gain Ratio Attribute Evaluation* (GRAE) avalia os atributos medindo a taxa de ganho em relação à classe (Han et al, 2011). Sua estrutura se dá por meio de uma árvore de decisão, onde os nós representam testes em relação a um ou mais atributos e as folhas, os resultados da decisão. A medida de ganho de informações é usada para selecionar os atributos de teste em cada nó da árvore de decisão de maiores valores. O seu funcionamento é explicado a seguir:

Seja S configurado com s amostras de dados e m classes distintas, a informação esperada para classificar uma amostra é fornecida por:

$$I(S) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (13)$$

em que p_i é a probabilidade de uma amostra arbitrária pertencer à classe C_i e ser estimada por s_i/s .

Seja o atributo A com v valores distintos e s_{ij} o número de amostras da classe C_i em um subconjunto S_j , S_j contém aquelas amostras em S que têm o valor a_j de A .

Assim, a entropia, ou informação esperada baseada na partição em subconjuntos por A , é dada por:

$$E(A) = - \sum_{i=1}^m I(S) \frac{s_{1i} + s_{2i} + \dots + s_{mi}}{s} \quad (14)$$

As informações de codificação obtidas pela ramificação de A são:

$$Gain(A) = I(S) - E(A) \quad (15)$$

A métrica de seleção de atributos *Gain Ratio Attribute Evaluation* usa a taxa de ganho que aplica normalização ao ganho de informação usando um valor definido como:

$$SplitInfo_A(S) = - \sum_{i=1}^v (|S_i|/|S|) \log_2(|S_i|/|S|) \quad (16)$$

O valor acima representa as informações geradas pela divisão do conjunto de dados de treinamento S em v partições correspondentes a v resultados de um teste no atributo A . A taxa de ganho é definida como:

$$Gain Ratio(A) = Gain(A)/SplitInfo_A(S) \quad (17)$$

O atributo com a maior taxa de ganho é selecionado como o atributo de divisão. O nó não folha da árvore de decisão gerada é considerado como atributos relevantes. Assim, o critério da taxa de ganho seleciona, dentre os atributos com ganho médio, ou melhor, o atributo que maximiza a taxa de ganho dividida por sua entropia (HALL, 1999).

Já o *Correlation Attribute Evaluation* (CorrAE) avalia o valor de um subconjunto de atributos considerando sua capacidade preditiva individual, juntamente com o grau de redundância entre eles (Hall e Holmes., 2003). É um algoritmo de filtro simples que classifica subconjuntos de atributos de acordo com uma função de avaliação heurística baseada em correlação. O viés da função de avaliação é direcionado a atributos altamente correlacionados com a classe e não correlacionados entre si. A função de

avaliação de subconjunto de atributos do *Correlation Attribute Evaluation* é mostrada na Equação 18 (HALL, 1999):

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{(k + k(k - 1)\overline{r_{ff}})}} \quad (18)$$

sendo M_S o “mérito” heurístico de um subconjunto de recursos S contendo k recursos, $\overline{r_{cf}}$ a correlação média entre a classe e o atributo ($f \in S$) e $\overline{r_{ff}}$ a intercorrelação média de atributos. O numerador da Equação 18 pode ser considerado uma indicação de quão preditivo da classe é um conjunto de atributos e o denominador de quanta redundância existe entre os recursos.

5.2.4.2 Segundo experimento

Conforme destacado anteriormente, no segundo experimento foram utilizadas as três cenas multitemporais: imagens de julho/2018 e janeiro e de abril de 2019. Como dito, a imagem de julho foi escolhida por oferecer condições desfavoráveis à discriminação de áreas de café infectado por nematoides. Os períodos de janeiro e abril podem oferecer condições mais favoráveis em relação ao mês de julho devido às condições ambientais e ciclos fenológicos envolvidos. Precipitações e temperaturas intermediárias ou moderadas podem trazer mais vigor às plantações de café comparando-se com períodos de seca e de baixas temperaturas, nos quais é possível observar desfolha mesmo em áreas saudáveis.

O mês de janeiro compreende as fases da vegetação e granação dos frutos, para o primeiro e segundo ano do ciclo fenológico do café, respectivamente. Além disso, o mês de janeiro possui precipitação moderada (média de 173,4 mm dos anos 2018 e 2019), em relação a novembro, mês de maior precipitação (338,3 mm). A temperatura no mês de janeiro também é moderada (24,8 °C) em relação aos meses mais frios e quentes de outros períodos.

Já a imagem de abril, foi escolhida por abranger as fases de indução e maturação das gemas florais (primeiro ano fenológico) e maturação dos frutos (segundo ano) do ciclo fenológico do café. Além disso, esse período possui precipitação intermediária (média de 73,6 mm), comparando-se com os meses de julho e janeiro e temperatura também moderada (média de 23,5 °C).

Assim, no segundo experimento as três imagens foram utilizadas com o objetivo de comparar os atributos extraídos para cada fase fenológica tratada. Para isso, inicialmente, foi utilizado o *software* ENVI 5.1 para a coleta das amostras de café infectado, café sadio e solo exposto a partir do segundo conjunto de dados de referência disponibilizado pela UFU e abrangendo a região mapeada pela Emater-MG. Na Tabela 4 estão indicados o número de elementos amostrais adquiridos para cada classe, ressaltando que a menor quantidade de pixels obtidos para a classe solo exposto decorre da menor representatividade dessa classe nos talhões incluídos na máscara da cultura de café.

Tabela 4 – Número de pixels de treinamento amostrados por classe

Classe	Número de pixels de treinamento
Café infectado	13824
Café sadio	16599
Solo exposto	1140

A resposta dos pixels amostrados foi vinculada com as imagens de cada uma das datas trabalhadas por meio do *software* ENVI, posteriormente, esses dados foram exportados em formato ASCII e convertidos para ARFF.

Do mesmo modo que no experimento 1, os arquivos ARFFs foram utilizados como dados de entrada para as métricas de seleção de atributos no ambiente WEKA. Foram aplicadas as mesmas métricas: *Classifier Attribute Evaluation*, *Gain Ratio Attribute Evaluation* e *Correlation Attribute Evaluation*.

5.2.5 Classificação

Com base nos resultados obtidos na seleção dos atributos que melhor separam as classes de interesse nos dois experimentos, foram definidos cinco conjuntos de atributos. Para cada conjunto de atributo, foram criados três arquivos ARFFs, um para cada data de aquisição da imagem (julho, janeiro e abril), em um total de 15 arquivos. Esses arquivos foram divididos de forma randômica (por meio de ferramenta disponível no WEKA) em dados de treinamento (80% dos pixels amostrados) e validação. Os arquivos de treinamento foram utilizados como dados de entrada para o treinamento dos algoritmos RNA e FA.

O treinamento da RNA consistiu na aplicação de arquiteturas de rede distintas aos diferentes conjuntos de atributos, a fim de selecionar a arquitetura que apresenta um

melhor desempenho na discriminação de áreas de café infectado. Os dados de treinamento e validação foram submetidos a uma configuração de rede neural constituída por camada de entrada, camadas escondidas e camada de saída, designada no *Multilayer Perceptron* (ML) presente no ambiente WEKA, assim como a definição dos parâmetros de configuração da rede estabelecendo o número de camadas escondidas, bem como a quantidade de elementos de processamento em cada camada. Para o algoritmo FA, somente foram definidos os parâmetros de processamento para o treinamento.

Em todos os treinamentos, somente as combinações de planos de entrada variaram, mantendo os mesmos parâmetros da RNA, desde número de camadas intermediárias, taxa de aprendizado, taxa de momento e ciclos de treinamento, e para o FA, o número de árvores (parâmetro *ntree*) e o número de variáveis amostradas aleatoriamente como candidatas em cada divisão (parâmetro *mtry*). Os arquivos de treinamento foram divididos, tanto pra RNA, quanto para FA, de modo que 80% dos pixels amostrados fossem utilizados para treinamento e 20% para validação, definidos de forma randômica pelo WEKA.

Para todos os treinamentos realizados foram definidos os seguintes parâmetros. No caso do algoritmo RNA adotou-se uma taxa de aprendizado (α) de 0,01, termo de momento 0,1, função de ativação sigmodal e número de ciclos de treinamento igual a 3000. O número de épocas de treinamento pode afetar consideravelmente a qualidade da resposta da rede. Com um número pequeno, a rede pode não extrair adequadamente as características dos dados de entrada, já um número elevado pode provocar um “super ajuste” dos pesos da rede, quando ela passa a memorizar os padrões fornecidos e se torna incapaz de generalizar as características extraídas para outros conjuntos de dados. Dessa forma, o número de épocas de treinamento (ciclos) foi definido a partir da análise da convergência do erro médio quadrático (EMQ) no treinamento, observando que, o EMQ alcançava valores estáveis em torno de 3000 iterações.

Para o algoritmo FA, o número de árvores foi estabelecido como sendo 1000. Lawrence et al. (2006) mostram que estudos anteriores comprovaram que não há aumento no erro além da criação de 1000 árvores de classificação a partir de amostras selecionadas aleatoriamente. A partir dos resultados de classificação utilizando FA relatados em estudos anteriores e revisados em Belgiu e Drăguț (2016) e Gislason et al. (2006), o parâmetro *mtry* foi estabelecido como a raiz quadrada do número de camadas disponíveis. Portanto, esse parâmetro foi calculado para cada conjunto de atributos.

Ao total, foram realizados e analisados 30 treinamentos, sendo 15 para cada um dos algoritmos com a finalidade de se identificar o conjunto de atributos e o algoritmo mais adequado para classificar os dados selecionados.

Com os resultados obtidos nos treinamentos, foi definido o conjunto de atributos mais adequado para a aplicação na classificação das imagens de julho, janeiro e abril, tanto para a RNA, quanto para o FA.

Para a classificação das cenas tomadas em cada data, foi gerada uma máscara contendo as áreas de café assistidas pela Emater-MG, para que o processo fosse realizado somente dentro dos talhões dessa cultura. Assim, três novos arquivos ARFFs foram elaborados, um para cada data, referente ao conjunto de atributos escolhido para a classificação. Nesse arquivo, os pixels devem ser rotulados como *unclassified* ou, simplesmente, sem rótulo (?), conforme exemplificado na Figura 24.

Figura 24 – Exemplo de um arquivo formato ARFF a ser classificado

```
@relation Amostras

@attribute classe {cafe_infectado,cafe_sadio,solo_exposto}
@attribute B4 NUMERIC
@attribute B7 NUMERIC
@attribute B8 NUMERIC
@attribute B11 NUMERIC
@attribute B12 NUMERIC
@attribute MSI_RE3 NUMERIC
@attribute TCARI NUMERIC
@attribute MCARI NUMERIC
@attribute CSI_RE3 NUMERIC
@attribute B8_glcm NUMERIC
@attribute B11_glcm NUMERIC

@data
?,0.0686,0.2741,0.3060,0.2055,0.1335,0.7497,-0.1655,-0.0552,2.0532,6.1241,4.4221
?,0.0456,0.2741,0.2851,0.2055,0.1335,0.7497,-0.1705,-0.0568,2.0532,5.9276,4.5081
?,0.0644,0.3027,0.2912,0.2053,0.1179,0.6782,-0.1648,-0.0549,2.5674,5.4792,4.1508
?,0.0675,0.3027,0.2912,0.2053,0.1179,0.6782,-0.1639,-0.0546,2.5674,5.0517,4.0031
?,0.0626,0.3488,0.3159,0.1976,0.1077,0.5665,-0.1513,-0.0504,3.2386,4.8659,3.7740
?,0.0542,0.2741,0.2853,0.2055,0.1335,0.7497,-0.1696,-0.0565,2.0532,6.0651,4.5765
?,0.0329,0.2622,0.2810,0.1825,0.1022,0.6960,-0.1428,-0.0476,2.5656,5.7351,4.5041
?,0.0277,0.2622,0.2751,0.1825,0.1022,0.6960,-0.1433,-0.0478,2.5656,5.5829,4.4341
?,0.0265,0.2662,0.2789,0.1811,0.0988,0.6803,-0.1444,-0.0481,2.6943,5.1639,4.1079
```

A classificação foi realizada no software WEKA, por meio da ferramenta *Supplied test set*. Essa ferramenta realiza a classificação do conjunto de dados a partir dos modelos da rede neural e de florestas aleatórias geradas no processo de treinamento.

5.2.5.1 Geração dos mapas resultantes da classificação

O produto de saída da classificação do WEKA indica as classes de saída previstas para cada pixel da cena classificada e respectivo erro de predição (conforme arquivo exemplificado na Figura 25), esse último associado à probabilidade de determinado pixel ter sido classificado corretamente. No entanto, para geração do mapa final da classificação, é necessário associar cada pixel classificado à sua posição na imagem, já que o software não gera a imagem da classificação.

Figura 25 – Exemplo do arquivo de saída da classificação do WEKA

inst#	actual	predicted	error	prediction
1	1:?	1:cafe_infectado		0.601
2	1:?	1:cafe_infectado		0.614
3	1:?	1:cafe_infectado		0.573
4	1:?	1:cafe_infectado		0.673
5	1:?	1:cafe_infectado		0.743
6	1:?	1:cafe_infectado		0.543
7	1:?	1:cafe_infectado		0.661
8	1:?	1:cafe_infectado		0.997
9	1:?	1:cafe_infectado		1
10	1:?	1:cafe_infectado		0.923
11	1:?	1:cafe_infectado		0.851
12	1:?	1:cafe_infectado		0.582
13	1:?	1:cafe_infectado		0.914
14	1:?	1:cafe_infectado		0.46
15	1:?	1:cafe_infectado		0.853
16	1:?	1:cafe_infectado		0.802
17	1:?	1:cafe_infectado		0.965

Utilizando o software ArcGis, o arquivo de saída da classificação foi convertido em formato imagem (*raster*) por meio da ferramenta *Point to Raster*. Esse processo foi realizado para os resultados referentes aos dois classificadores, para cada uma das datas. No total foram gerados seis mapas temáticos mostrando a distribuição espacial das classes de saída.

5.2.5.2 Validação dos dados classificados

O desempenho dos classificadores RNA e FA foi analisado a partir dos dados de validação, dados esses não utilizados na construção dos modelos de treinamento, utilizando métricas clássicas geradas a partir da matriz de confusão. As métricas calculadas foram: Acurácia Global (AG), Acurácia do Produtor (AP), Acurácia do Usuário (AU), coeficiente *Kappa* e *F-score*.

A matriz de confusão é expressa por meio da correlação de dados de referência, nesse caso representados pelos dados de validação, com os dados classificados. A AG é uma medida simples, calculada pela soma do total dos pixels corretamente classificados dividida pelo número total de pixels da matriz de confusão, o que corresponde à porcentagem de elementos corretamente classificados. A AP indica a probabilidade de um pixel de referência ter sido corretamente classificado de acordo com os dados de referência. Já a AU, indica a probabilidade de uma categoria classificada representar a mesma categoria no campo (CONGALTON, 1991).

O coeficiente *Kappa* também é um dos indicadores de qualidade que podem ser quantificados após a construção da matriz de confusão, sendo um índice que retrata estatisticamente o grau de concordância entre os dados de referência e os dados classificados (CONGALTON e GREEN, 2008).

A análise das previsões dos classificadores também foi realizada por meio das medidas de *Precision* (confiança), *Recall* (sensibilidade) e *F-Score*. *Precision* descreve a proporção de casos positivos identificados corretamente ou não. *Recall* é a proporção de casos positivos e negativos identificados corretamente. A medida *F-Score* é a média harmônica de *Precision* e *Recall*, onde uma pontuação *F-Score* atinge seu melhor valor em 1 (*Precision* e *Recall* perfeitos) (POWERS, 2008). As equações 19, 20 e 21 mostram como as medidas são calculadas, respectivamente.

$$Precision = \frac{VP}{VP + FP} \quad (19)$$

$$Recall = \frac{VP}{VP + FN} \quad (20)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (21)$$

em que VP (verdadeiro positivo) é o número de elementos identificados na classe de interesse coincidente com os dados de referência; FP (falso positivo) é o número de pixels identificados na classe de interesse que não são coincidentes com os dados de referência e FN o número de elementos da outra classe que não são coincidente com os dados de referência.

Além disso, a partir dos valores de probabilidade computados pelo próprio WEKA e disponibilizados no produto de saída da classificação, foram geradas imagens de incerteza para cada um dos classificadores em cada uma das datas, para a análise da incerteza das cenas classificadas. A incerteza para cada pixel da imagem classificada é dada a partir de um valor de probabilidade, que indica a chance de o pixel ter sido classificado corretamente. Esse valor representa a probabilidade *a posteriori* de um determinado pixel da imagem pertencer a uma das classes de saída (RICHARDS e JIA, 2006). Para ter acesso à incerteza, basta subtrair o valor de probabilidade de 1.

6. RESULTADOS

6.1 Análise de atributos

Os resultados da análise de atributos desenvolvida a partir dos dois experimentos descritos na seção anterior são apresentados nas subseções a seguir.

6.1.1 Primeiro experimento

Para mostrar os resultados do primeiro experimento foram elaborados gráficos indicando a média dos valores de reflectância de superfície dos pixels, a partir das amostras coletadas de café infectado e café sadio para cada atributo (bandas e índices) na imagem de julho de 2018, assim como o valor do índice M de separabilidade para os diferentes índices de vegetação. Para analisar a importância dos diferentes atributos na discriminação do café infectado do sadio, foram utilizadas as métricas as *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)*.

Com base nos valores médios de reflectância de superfície das bandas espectrais, a Figura 26 mostra as diferenças entre as médias de amostras de café infectado e sadio, referente à cena de julho de 2018, enquanto que os gráficos da Figura 27 indicam a variação espectral das bandas, por meio de suas médias e desvios padrão, para cada classe individualmente. A Tabela 5 apresenta os valores das médias e respectivos desvios-padrão da reflectância de superfície obtidos a partir das amostras de café infectado e sadio, bem como a diferença absoluta entre as médias, para cada banda espectral do Sentinel-2.

Figura 26 – Médias dos valores de reflectância de superfície dos pixels amostrados de café infectado e sadio obtidas das bandas espectrais do Sentinel-2 para a cena de julho de 2018

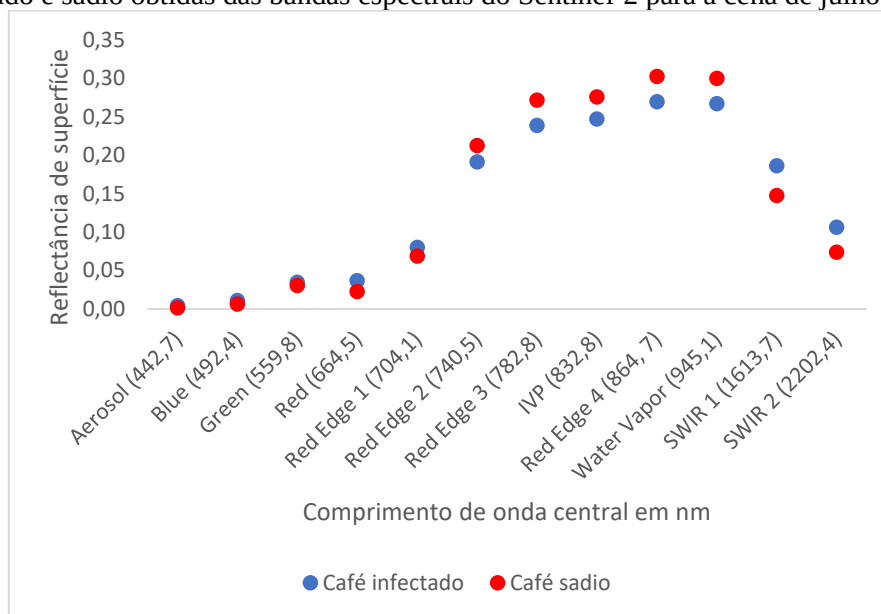


Figura 27 – Variação espectral nas médias dos valores de reflectância de superfície e seus respectivos desvios-padrão, obtidos para as amostras de café infectado (a) e sadio (b) nas bandas espectrais da cena de julho de 2018 do Sentinel-2

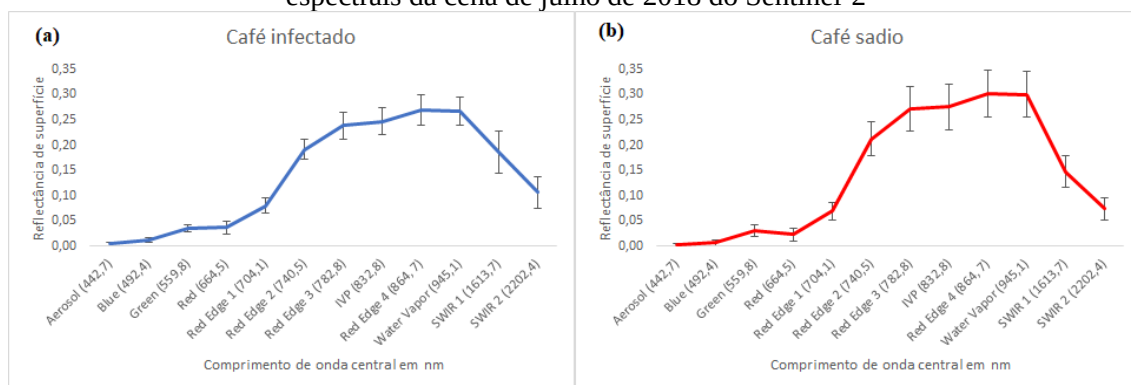


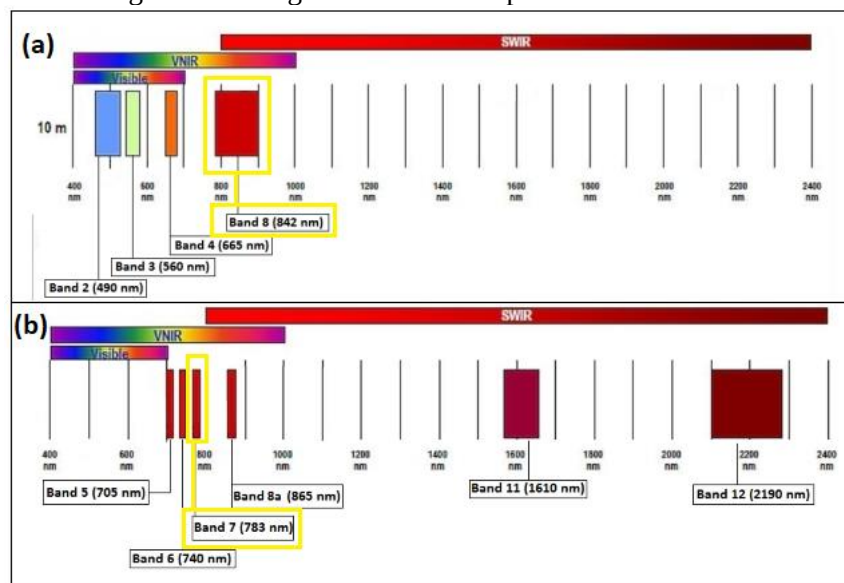
Tabela 5: Médias e desvios-padrão da reflectância de superfície obtidos a partir das amostras de café infectado e sadio e a diferença absoluta entre as médias, para cada banda espectral do Sentinel-2

Região espectral	Café infectado		Café Sadio		Diferença das médias (nm)
	Média (nm)	Desvio Padrão (nm)	Média (nm)	Desvio Padrão (nm)	
Aerossol	0,0045	0,0035	0,0018	0,0033	0,0027
Azul	0,0111	0,0046	0,0067	0,0060	0,0044
Verde	0,0347	0,0064	0,0306	0,0117	0,0042
Vermelho	0,0370	0,0124	0,0226	0,0119	0,0144
Red-Edge 1	0,0804	0,0143	0,0691	0,0181	0,0114
Red-Edge 2	0,1914	0,0200	0,2127	0,0343	-0,0213
Red-Edge	0,2387	0,0267	0,2717	0,0437	-0,0330
IVP	0,2471	0,0270	0,2758	0,0453	-0,0287
Red-Edge 4	0,2696	0,0298	0,3023	0,0466	-0,0327
Water vapor	0,2671	0,0271	0,2999	0,0454	-0,0328
SWIR 1	0,1864	0,0410	0,1475	0,0319	0,0389
SWIR 2	0,1067	0,0309	0,0741	0,0215	0,0326

De acordo com a Figura 26 e a Tabela 5, as bandas B1 (Aerossol), B2 (Azul) e B3 (Verde) não apresentaram diferença considerável entre os valores avaliados (0,0027; 0,0044 e 0,0042 nm, respectivamente). Já a banda B4 (Vermelho), mostrou diferença significativa entre as médias dos valores de reflectância das classes (0,0144 nm). Como esperado, a banda do SWIR 1 (B11, infravermelho de ondas curtas 1) foi a que apresentou maior separabilidade entre as classes (0,0389 nm). Ressalta-se que essa região do espectro é sensível ao teor de umidade da planta e, portanto, essa discriminação pode ocorrer devido ao aumento na energia refletida em razão, principalmente, do baixo conteúdo de água presente na planta causado pelos nematoides.

A região do NIR (B8, 832,8 nm) também mostra uma diferença considerável nos valores de reflectância (-0,0287). Essa região é caracterizada pela alta reflectância da planta, porém o conteúdo de água presente na planta, entre outros fatores, também interfere na sua reflectância. Destaca-se também que a banda B7 (RE3, 782,8 nm) indicou maior separabilidade que a banda do NIR (infravermelho próximo), entre o café safio e infectado. Apesar de espectralmente próximas, essas duas bandas (NIR e RE3) definem larguras de faixas espectrais diferentes. O RE3 se refere a uma banda muito mais estreita (20 nm) que o NIR (106 nm), permitindo um maior detalhamento espectral dos alvos. A Figura 28 mostra as larguras das bandas espectrais do Sentinel-2 com (a) 10 metros de resolução espacial; e (b) 20 metros de resolução espacial. Nessa figura, destacou-se em amarelo as bandas do NIR (Band 8) e RE3 (Band 7), em (a) e (b), respectivamente.

Figura 28 – Largura das bandas espectrais do Sentinel-2



Fonte: ESA, 2019

Barnes et al. (2000) realizaram um estudo com o objetivo principal de distinguir vegetação sob estresse de hídrico e deficiência no teor de nitrogênio, em uma área de cultivo de algodão na região de Maricopa (Arizona), com base em dados de sensoriamento remoto. Os autores mostram a importância das faixas espectrais do vermelho e infravermelho próximo no monitoramento da vegetação e destacaram a faixa do vermelho limítrofe, que sintetiza a forte absorção da luz vermelha pela clorofila e os comprimentos de onda do infravermelho próximo altamente reflexivos. A partir dessas observações, os autores propuseram a otimização de alguns índices de vegetação, como por exemplo, o NDVI, substituindo a região do infravermelho próximo pelo *Red Edge* (~790 nm).

Com base no estudo desenvolvido por Barnes et al. (2000) e na maior diferença entre as médias apresentada pela banda do RE3 em relação a do NIR, optou-se por, além de gerar os índices de vegetação na sua formulação original, substituir a banda do NIR pela banda RE3 nos índices que utilizam a banda espectral do NIR. A Figura 29 e a Figura 30 mostram as diferenças entre as médias dos índices de vegetação obtidos das amostras de café infectado e sadio, por meio de gráficos desses valores normalizados e não normalizados, respectivamente. Os índices gerados que tiveram a substituição da banda do NIR pela banda do RE3 foram: NDVI, NDWI, MSI e CSI e são apresentados nos gráficos como: NDVI_RE3, NDWI_RE3, MSI_RE3 e CSI_RE3. Em seguida, a Tabela 6 indica as médias, desvios-padrão e diferença entre as médias de cada classe.

Figura 29 – Médias dos índices de vegetação normalizados para amostras de café infectado e sadio, obtidas a partir da cena de julho de 2018 do Sentinel-2

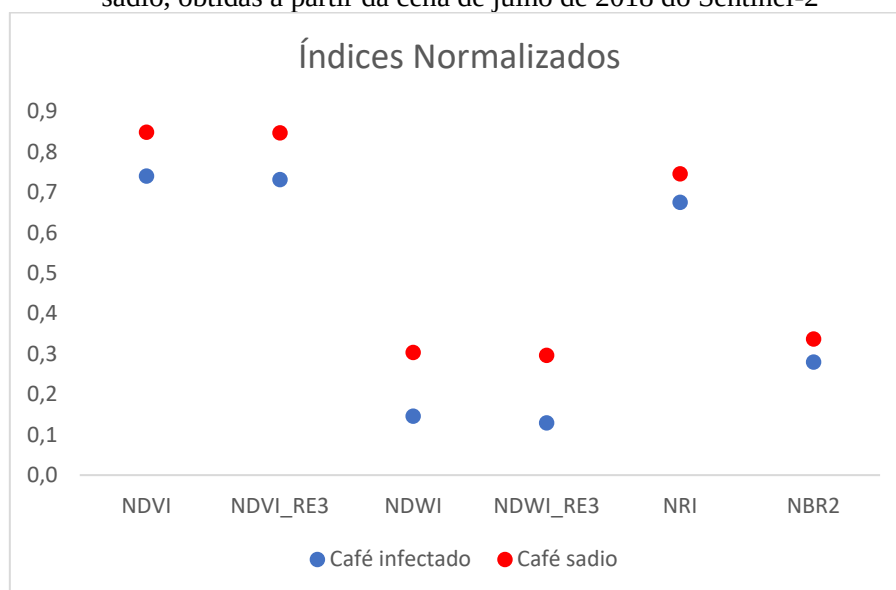


Figura 30 – Médias dos índices de vegetação não-normalizados para amostras de café infectado e sadio, obtidas a partir da cena de julho de 2018 do Sentinel-2

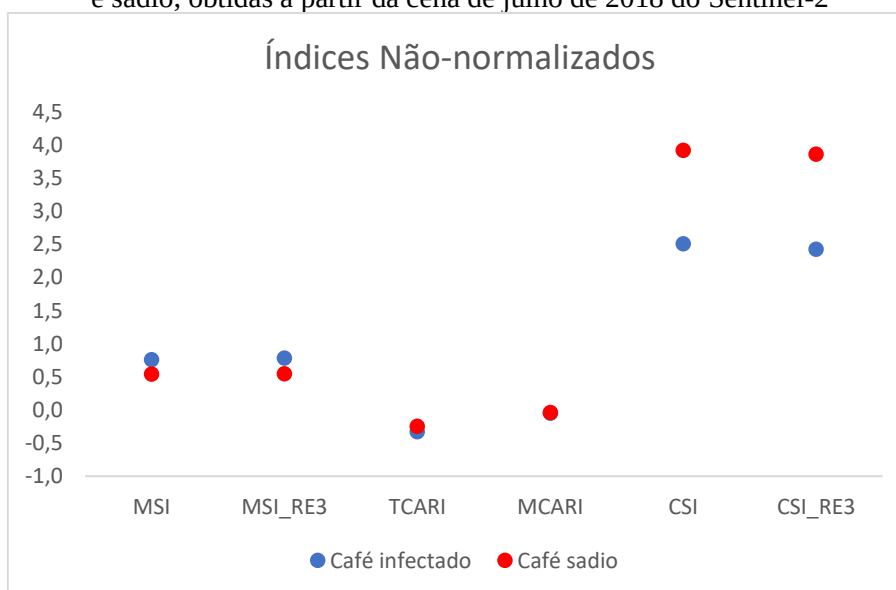


Tabela 6: Médias e desvios-padrão dos valores de índices de vegetação obtidos a partir das amostras de café infectado e sadio e a diferença absoluta entre as médias

Índices	Café infectado		Café sadio		Diferença das médias	
	Média	Desvio Padrão	Média	Desvio Padrão		
NDVI	0,7401	0,0782	0,8488	0,0671	-0,1087	Normalizados
NDVI_RE3	0,7319	0,0811	0,8468	0,0679	-0,1149	
NDWI	0,1465	0,1026	0,3035	0,0847	-0,1570	
NDWI_RE3	0,1295	0,1061	0,2970	0,0860	-0,1675	
NRI	0,6749	0,0570	0,7459	0,0598	-0,0710	
NBR2	0,2799	0,0405	0,3373	0,0371	-0,0574	
MSI	0,7581	0,1538	0,5413	0,1077	0,2168	Não Normalizados
MSI_RE3	0,7860	0,1630	0,5493	0,1097	0,2367	
TCARI	-0,3299	0,0853	-0,2462	0,0512	-0,0837	
MCARI	-0,0495	0,0099	-0,0401	0,0086	-0,0094	
CSI	2,5108	0,8182	3,9189	0,9167	-1,4080	
CSI_RE3	2,4293	0,8097	3,8654	0,9152	-1,4361	

Dentre os índices normalizados, o que apresentou maior diferença entre as médias do café sadio e infectado foi o NDWI_RE3 (-0,1675), com destaque também para o NDVI_RE3 (-0,1149). Dos índices não normalizados, o CSI_RE3 apresentou o maior valor (-1,4361), porém resultou nos maiores valores de desvio padrão para as classes de café infectado e sadio, abaixo apenas do CSI.

De acordo com os resultados, todos os índices em que foi feita a substituição da banda do NIR pelo RE3, definiram uma maior diferença entre as médias das classes de café infectado e sadio, em relação aos índices nas suas formulações originais. Os desvios-padrão, em ambos os casos, apresentaram valores semelhantes.

O índice M de separabilidade também foi aplicado para investigar os índices que melhor discriminam as classes, reforçando que valores de M maiores que 1 (um), indicam classes mais discrimináveis. Os valores de M obtidos para os índices de vegetação analisados são apresentados na Tabela 7:

Tabela 7: Valores M obtidos para os índices de vegetação normalizados e não-normalizados

Índice de Separabilidade M		
NDVI	0,7485	
NDVI_RE3	0,7710	
NDWI	0,8377	Normalizados
NDWI_RE3	0,8722	
NRI	0,6084	
NBR2	0,7395	
MSI	0,8290	
MSI_RE3	0,8676	
TCARI	0,6131	Não- Normalizados
MCARI	0,5100	
CSI	0,8115	
CSI_RE3	0,8326	

Nenhum dos índices apresentou valor maior que 1, contudo, os índices normalizados NDWI_RE3 (0,8722), NDWI (0,8377), e NDVI_RE3 (0,7710), e os não-normalizados MSI_RE3 (0,8676), CSI_RE3 (0,8326), MSI (0,8290) e CSI (0,8115) resultaram nos maiores valores.

Em seguida, foram aplicadas as métricas de seleção de atributos *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)*, com o objetivo de investigar os atributos de maior importância para a discriminação das classes, agora com base nas amostras de café infectado, café sadio e solo exposto, extraídos de uma pequena porção da área de estudo (área teste definida na Figura 22). As métricas indicam a contribuição de cada atributo na discriminação entre as classes, expressas na forma de um *ranking* de importância, conforme mostrado na Tabela 8. Para esse processamento, apenas as bandas espectrais do SWIR, NIR, RE3 e Vermelho foram consideradas.

Tabela 8: *Ranking* de Importância dos atributos definido pelas métricas *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)*

ClassAE		GRAE		CorrAE	
CSI_RE3	0,39	CSI	0,4	RE3	0,512
MSI_RE3	0,35	NDVI_RE3	0,38	NIR	0,493
NDWI_RE3	0,34	NDVI	0,36	MCARI	0,386
NBR2	0,33	MSI_RE3	0,35	SWIR 1	0,379
NDVI_RE3	0,27	NDWI_RE3	0,35	CSI_RE3	0,375
SWIR 2	0,27	CSI_RE3	0,33	SWIR 2	0,367
NRI	0,26	MSI	0,31	CSI	0,367
SWIR 1	0,26	NDWI	0,31	TCARI	0,36
NDVI	0,25	NBR2	0,3	Red	0,358
RE3	0,24	Red	0,3	MSI	0,331
CSI	0,24	NRI	0,3	MSI_RE3	0,326
NDWI	0,21	TCARI	0,28	NBR2	0,322
MSI	0,21	SWIR 2	0,26	NDVI	0,315
MCARI	0,21	RE3	0,22	NDVI_RE3	0,311
TCARI	0,2	SWIR 1	0,22	NRI	0,306
Red	0,2	MCARI	0,22	NDWI	0,279
NIR	0,18	NIR	0,19	NDWI_RE3	0,271

As métricas *ClassAE* e *GRAE* apontam os índices CSI_RE3, MSI_RE3, NDWI_RE3 e NDVI_RE3 entre os seis atributos mais importantes. Pode-se verificar pela Tabela 8 que os atributos que apresentaram maior importância para ambas as métricas foram aqueles gerados a partir da banda do vermelho limítrofe, RE3. Dentre as seis características que apresentaram maior importância, o índice CSI_RE3 foi o único apresentado nas três métricas.

Segundo Hall (1999), a importância é direcionada a subconjuntos de atributos que são altamente correlacionados com a classe e não correlacionados entre si. Por outro lado, uma alta intercorrelação entre os atributos, diminui essa importância. Com base nisso, para avaliar a correlação entre os atributos, a matriz de correlação foi calculada e é apresentada na Figura 31.

Figura 31 – Matriz de correlação dos atributos

	B4	B7	B8	B11	B12	CSI_RE3	CSI	MCARI	MSI	MSI_RE3	NBR2	NDVI	NDVI_RE3	NDWI	NDWI_RE3	NRI	TCARI
B4	1.000	-0.431	-0.438	0.979	0.987	-0.770	-0.780	-0.980	0.965	0.961	-0.888	-0.983	-0.982	-0.942	-0.936	-0.920	-0.966
B7	-0.431	1.000	0.989	-0.348	-0.402	0.766	0.762	0.335	-0.570	-0.579	0.632	0.564	0.571	0.652	0.663	0.613	0.391
B8	-0.438	0.989	1.000	-0.350	-0.406	0.750	0.753	0.338	-0.576	-0.577	0.628	0.572	0.574	0.654	0.657	0.619	0.388
B11	0.979	-0.348	-0.350	1.000	0.993	-0.761	-0.770	-0.998	0.955	0.952	-0.867	-0.951	-0.950	-0.932	-0.927	-0.854	-0.992
B12	0.987	-0.402	-0.406	0.993	1.000	-0.784	-0.792	-0.993	0.964	0.960	-0.904	-0.967	-0.965	-0.946	-0.940	-0.881	-0.984
CSI_RE3	-0.770	0.766	0.750	-0.761	-0.784	1.000	0.999	0.744	-0.823	-0.826	0.928	0.830	0.834	0.913	0.920	0.864	0.795
CSI	-0.780	0.762	0.753	-0.770	-0.792	0.999	1.000	0.753	-0.832	-0.833	0.932	0.840	0.842	0.920	0.924	0.872	0.802
MCARI	-0.980	0.335	0.338	-0.998	-0.993	0.744	0.753	1.000	-0.952	-0.948	0.859	0.949	0.947	0.924	0.918	0.850	0.984
MSI	0.965	-0.570	-0.576	0.955	0.964	-0.823	-0.832	-0.952	1.000	0.998	-0.898	-0.986	-0.985	-0.981	-0.976	-0.893	-0.960
MSI_RE3	0.961	-0.579	-0.577	0.952	0.960	-0.826	-0.833	-0.948	0.998	1.000	-0.896	-0.983	-0.985	-0.979	-0.978	-0.890	-0.959
NBR2	-0.888	0.632	0.628	-0.867	-0.904	0.928	0.932	0.859	-0.898	-0.896	1.000	0.921	0.922	0.945	0.946	0.926	0.878
NDVI	-0.983	0.564	0.572	-0.951	-0.967	0.830	0.840	0.949	-0.986	-0.983	0.921	1.000	0.999	0.976	0.971	0.950	0.949
NDVI_RE3	-0.982	0.571	0.574	-0.950	-0.965	0.834	0.842	0.947	-0.985	-0.985	0.922	0.999	1.000	0.976	0.974	0.950	0.949
NDWI	-0.942	0.652	0.654	-0.932	-0.946	0.913	0.920	0.924	-0.981	-0.979	0.945	0.976	0.976	1.000	0.998	0.924	0.944
NDWI_RE3	-0.936	0.663	0.657	-0.927	-0.940	0.920	0.924	0.918	-0.976	-0.978	0.946	0.971	0.974	0.998	1.000	0.921	0.942
NRI	-0.920	0.613	0.619	-0.854	-0.881	0.864	0.872	0.850	-0.893	-0.890	0.926	0.950	0.950	0.924	0.921	1.000	0.854
TCARI	-0.966	0.391	0.388	-0.992	-0.984	0.795	0.802	0.984	-0.960	-0.959	0.878	0.949	0.949	0.944	0.942	0.854	1.000

Correlação é uma medida da relação estatística entre duas variáveis. Quanto maior o seu valor, maior a redundância espectral entre as variáveis. Como esperado, os índices de vegetação NDVI, NDWI, CSI e MSI correlacionaram-se com valor próximo a 1, em relação aos índices NDVI_RE3, NDWI_RE3, CSI_RE3 e MSI_RE3, respectivamente. De maneira geral, os índices de vegetação que apresentaram uma correlação um pouco menor com os demais atributos foram CSI_RE3, CSI e MCARI. As altas correlações podem ser explicadas pelo uso de basicamente as mesmas bandas na formulação dos índices. Em relação às bandas espectrais, em geral, aquelas que apresentaram os menores valores de correlação foram a RE3 e NIR.

A métrica *CorrAE*, diferentemente das demais métricas aplicadas, atribui pouca importância aos índices MSI_RE3, NDWI_RE3 e NDVI_RE3 na separação entre as classes, provavelmente decorrente das altas intercorrelações apresentadas entre eles. Por outro lado, as bandas espectrais do RE3 (B7) e NIR (B8), apesar de correlacionadas, definiram maior importância para a previsão da classe de interesse.

6.1.2 Segundo experimento

O segundo experimento também foi desenvolvido a fim de orientar a seleção de um conjunto de atributos que melhor discrimine áreas de café infectado por nematoides, porém incorporando atributos texturais e um delineamento amostral que viabilize a análise de imagens do Sentinel-2 tomadas em fases representativas do ciclo de

desenvolvimento do café. Nesse sentido, foram utilizadas imagens de julho/2018, janeiro/2019 e abril/2019 e foram coletadas novas amostras com base no segundo conjunto de dados de referência disponibilizados pela Universidade Federal de Uberlândia (UFU).

Os valores de importância dos atributos resultantes das métricas de seleção, assim como a matriz de correlação dos atributos referente à cena tomada no mês de janeiro/2019, podem ser observados na Tabela 9 e na Figura 32, respectivamente.

Tabela 9 - Ranking de Importância dos atributos definido pelas métricas *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)* para a imagem de janeiro/2019

<i>ClassAE</i>		<i>GRAE</i>		<i>CorrAE</i>	
CSI_RE3	0,4173	VERMELHO	0,2049	CSI	0,5226
MSI_RE3	0,3046	NDVI	0,1917	CSI_RE3	0,5185
NDWI_RE3	0,2877	NDVI_RE3	0,1896	SWIR1	0,5158
SWIR2	0,2692	NRI	0,16418	MCARI	0,5092
NBR2	0,2596	CSI	0,16232	TCARI	0,5092
SWIR1	0,2503	MSI	0,14619	SWIR2	0,4836
VERMELHO	0,2478	CSI_RE3	0,14616	NBR2	0,4831
MCARI	0,2266	NBR2	0,14523	NDWI_RE3	0,4416
NRI	0,2186	NDWI	0,14488	NDWI	0,4399
TCARI	0,2179	MSI_RE3	0,13842	NRI	0,4263
NDVI_RE3	0,2036	SWIR2	0,13542	VERMELHO	0,3973
NDVI	0,2015	NDWI_RE3	0,1343	MSI_RE3	0,3881
CSI	0,1924	SWIR1	0,12181	MSI	0,377
NDWI	0,1919	MCARI	0,11373	NDVI_RE3	0,3605
MSI	0,1865	TCARI	0,10954	NDVI	0,3532
RE3	0,1485	NIR	0,09861	NIR	0,1862
NIR	0,1074	RE3	0,07987	RE3	0,1815
SWIR1_glcm	0,0161	SWIR1_glcm	0,02338	SWIR1_glcm	0,037
NIR_glcm	-0,0275	NIR_glcm	0,00922	NIR_glcm	0,0138

Figura 32 – Matriz de correlação dos atributos referente ao mês de janeiro/2019

	B4	B7	B8	B11	B12	NDVI	NDVI_RE3	NDWI	NDWI_RE3	NRI	NBR2	MSI	MSI_RE3	TCARI	MCARI	CSI	CSI_RE3	B8_glcm	B11_glcm
B4	1,000	-0,738	-0,760	0,830	0,942	-0,977	-0,978	-0,939	-0,925	-0,965	-0,926	0,955	0,946	-0,840	-0,840	-0,835	-0,828	0,190	0,098
B7	-0,738	1,000	0,989	-0,489	-0,675	0,819	0,824	0,817	0,817	0,793	0,793	-0,824	-0,828	0,490	0,490	0,729	0,731	-0,404	-0,182
B8	-0,760	0,989	1,000	-0,480	-0,677	0,844	0,844	0,820	0,809	0,810	0,794	-0,838	-0,831	0,485	0,485	0,716	0,712	-0,377	-0,177
B11	0,830	-0,489	-0,480	1,000	0,955	-0,747	-0,761	-0,885	-0,893	-0,772	-0,847	0,829	0,852	-0,993	-0,993	-0,890	-0,888	0,212	0,083
B12	0,942	-0,675	-0,677	0,955	1,000	-0,899	-0,908	-0,966	-0,965	-0,900	-0,948	0,944	0,954	-0,957	-0,957	-0,917	-0,914	0,221	0,091
NDVI	-0,977	0,819	0,844	-0,747	-0,899	1,000	0,999	0,940	0,922	0,971	0,929	-0,971	-0,957	0,760	0,760	0,813	0,806	-0,172	-0,102
NDVI_RE3	-0,978	0,824	0,844	-0,761	-0,908	0,999	1,000	0,948	0,933	0,973	0,937	-0,974	-0,964	0,774	0,774	0,826	0,820	-0,183	-0,105
NDWI	-0,939	0,817	0,820	-0,885	-0,966	0,940	0,948	1,000	0,997	0,936	0,969	-0,980	-0,987	0,884	0,884	0,943	0,939	-0,278	-0,144
NDWI_RE3	-0,925	0,817	0,809	-0,893	-0,965	0,922	0,933	0,997	1,000	0,924	0,967	-0,969	-0,983	0,891	0,891	0,951	0,950	-0,294	-0,147
NRI	-0,965	0,793	0,810	-0,772	-0,900	0,971	0,973	0,936	0,924	1,000	0,956	-0,928	-0,922	0,773	0,773	0,877	0,870	-0,208	-0,150
NBR2	-0,926	0,793	0,794	-0,847	-0,948	0,929	0,937	0,969	0,967	0,956	1,000	-0,937	-0,944	0,840	0,840	0,951	0,948	-0,244	-0,162
MSI	-0,955	-0,824	-0,838	0,829	0,944	-0,971	-0,974	-0,980	-0,969	-0,928	-0,937	1,000	0,996	-0,842	-0,842	-0,865	-0,859	0,218	0,101
MSI_RE3	0,946	-0,828	-0,831	0,852	0,954	-0,957	-0,964	-0,987	-0,983	-0,922	-0,944	0,996	1,000	-0,862	-0,862	-0,886	-0,883	0,242	0,106
TCARI	-0,840	0,490	0,485	-0,993	-0,957	0,760	0,774	0,884	0,891	0,773	0,840	-0,842	-0,862	1,000	1,000	0,870	0,867	-0,197	-0,062
MCARI	-0,840	0,490	0,485	-0,993	-0,957	0,760	0,774	0,884	0,891	0,773	0,840	-0,842	-0,862	1,000	1,000	0,870	0,867	-0,197	-0,062
CSI	-0,835	0,729	0,716	-0,890	-0,917	0,813	0,826	0,943	0,951	0,877	0,951	-0,865	-0,886	0,870	0,870	1,000	0,999	-0,325	-0,218
CSI_RE3	-0,828	0,731	0,712	-0,888	-0,914	0,806	0,820	0,939	0,950	0,870	0,948	-0,859	-0,883	0,867	0,867	0,999	1,000	-0,330	-0,218
B8_glcm	0,190	-0,404	-0,377	0,212	0,221	-0,172	-0,183	-0,278	-0,294	-0,208	-0,244	0,218	0,242	-0,197	-0,197	-0,325	-0,330	1,000	0,553
B11_glcm	0,098	-0,182	-0,177	0,083	0,091	-0,102	-0,105	-0,144	-0,147	-0,150	-0,162	0,101	0,106	-0,062	-0,062	-0,218	-0,218	0,553	1,000

As métricas *ClassAE* e *CorrAE* apresentaram as bandas do SWIR 1 e do SWIR 2 entre os seis atributos de maior importância. O índice NBR2 também obteve posição significativa no *ranking* de importância das três métricas, assim como o índice CSI_RE3 que aparece entre os sete atributos de maior importância nas três métricas aplicadas para a imagem de janeiro/2019. As três métricas também tiveram comportamento semelhante para os atributos de menor importância, apresentando as imagens de textura (SWIR1_glcm e NIR_glcm) e as bandas do NIR e RE3 nas últimas posições do *ranking*.

De acordo com a Figura 32, as imagens de textura obtidas a partir das imagens SWIR1 e NIR do Sentinel-2 foram as que apresentaram as menores correlações em relação aos demais atributos, seguidas das bandas RE3 e NIR. Os demais atributos mostraram-se fortemente correlacionados. A banda B11 obteve baixa correlação com as bandas RE3 e NIR e imagens de textura, e alta correlação em relação às demais bandas e índices.

Apesar das imagens de textura SWIR1_glcm e NIR_glcm e das bandas RE3 e NIR serem os atributos que apresentaram menor correlação em relação aos demais atributos, a métrica *CorrAE* os apresentou como os atributos de menor importância, coincidentemente com as demais métricas. Nesse caso, esses atributos texturais não foram significativos na discriminação entre as classes.

Os valores de importância dos atributos resultantes das métricas de seleção e respectiva matriz de correlação dos atributos, obtidos a partir da imagem de abril de 2019 são apresentados na Tabela 10 e Figura 33, respectivamente.

Tabela 10 - *Ranking* de Importância dos atributos definido pelas métricas *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)* para a imagem de abril/2019

<i>ClassAE</i>		<i>GRAE</i>		<i>CorrAE</i>	
CSI_RE3	0,4301	NDVI	0,1794	CSI_RE3	0,622
NBR2	0,3317	VERMELHO	0,1736	CSI	0,621
MSI_RE3	0,3315	NRI	0,1729	NBR2	0,582
NDWI_RE3	0,3253	NDVI_RE3	0,1729	NDWI_RE3	0,535
VERMELHO	0,3166	CSI	0,1618	NDWI	0,521
SWIR2	0,3089	NBR2	0,1614	NRI	0,507
NRI	0,3089	CSI_RE3	0,1568	SWIR2	0,471
NDVI	0,3044	MSI	0,1461	MSI_RE3	0,461
NDVI_RE3	0,303	NDWI	0,1452	MCARI	0,449
NDWI	0,2744	MSI_RE3	0,143	TCARI	0,449
MSI	0,2718	NDWI_RE3	0,1416	SWIR1	0,447
CSI	0,246	SWIR2	0,1326	MSI	0,441
MCARI	0,2457	NIR	0,1146	RE3	0,436
SWIR1	0,2444	TCARI	0,1123	NIR	0,425
TCARI	0,2409	RE3	0,1069	NDVI_RE3	0,418
RE3	0,2317	MCARI	0,1065	NDVI	0,403
NIR	0,1855	SWIR1	0,0971	VERMELHO	0,395
SWIR1_glcm	0,0421	SWIR1_glcm	0,0332	SWIR1_glcm	0,297
NIR_glcm	-0,0209	NIR_glcm	0,0259	NIR_glcm	0,129

Figura 33 – Matriz de correlação dos atributos referente ao mês de abril/2019

	Red	RE3	IVP	SWIR1	SWIR2	NDVI	NDVI_RE3	NDWI	NDWI_RE3	NRI	NBR2	MSI	MSI_RE3	TCARI	MCARI	CSI	CSI_RE3	IVP_glcm	SWIR1_glcm
Red	1,000	-0,577	-0,611	0,848	0,938	-0,984	-0,983	-0,907	-0,889	-0,947	-0,879	0,942	0,927	-0,882	-0,882	-0,767	-0,759	0,113	0,251
RE3	-0,577	1,000	0,990	-0,380	-0,542	0,649	0,660	0,738	0,749	0,696	0,707	-0,688	-0,702	0,400	0,400	0,768	0,774	-0,409	-0,333
IVP	-0,611	0,990	1,000	-0,391	-0,557	0,683	0,689	0,753	0,753	0,722	0,713	-0,710	-0,714	0,416	0,416	0,766	0,765	-0,393	-0,318
SWIR1	0,848	-0,380	-0,391	1,000	0,965	-0,806	-0,817	-0,887	-0,885	-0,776	-0,835	0,883	0,892	-0,987	-0,987	-0,793	-0,787	0,153	0,295
SWIR2	0,938	-0,542	-0,557	0,965	1,000	-0,919	-0,927	-0,952	-0,946	-0,891	-0,924	0,958	0,960	-0,971	-0,971	-0,850	-0,845	0,164	0,304
NDVI	-0,984	0,649	0,683	-0,806	-0,919	1,000	0,999	0,926	0,907	0,961	0,896	-0,959	-0,944	0,846	0,846	0,785	0,777	-0,108	-0,262
NDVI_RE3	-0,983	0,660	0,689	-0,817	-0,927	0,999	1,000	0,936	0,920	0,965	0,907	-0,964	-0,952	0,854	0,854	0,801	0,794	-0,115	-0,272
NDWI	-0,907	0,738	0,753	-0,887	-0,952	0,926	0,936	1,000	0,997	0,923	0,953	-0,978	-0,982	0,895	0,895	0,933	0,928	-0,250	-0,377
NDWI_RE3	-0,889	0,749	0,753	-0,885	-0,946	0,907	0,920	0,997	1,000	0,911	0,954	-0,967	-0,979	0,889	0,889	0,943	0,941	-0,264	-0,390
NRI	-0,947	0,696	0,722	-0,776	-0,891	0,961	0,965	0,923	0,911	1,000	0,945	-0,907	-0,899	0,808	0,808	0,869	0,862	-0,146	-0,338
NBR2	-0,879	0,707	0,713	-0,835	-0,924	0,896	0,907	0,953	0,954	0,945	1,000	-0,907	-0,915	0,842	0,842	0,954	0,951	-0,206	-0,416
MSI	0,942	-0,688	-0,710	0,883	0,958	-0,959	-0,964	-0,978	-0,967	-0,907	-0,907	1,000	0,996	-0,904	-0,904	-0,848	-0,841	0,188	0,295
MSI_RE3	0,927	-0,702	-0,714	0,892	0,960	-0,944	-0,952	-0,982	-0,979	-0,899	-0,915	0,996	1,000	-0,908	-0,908	-0,866	-0,862	0,204	0,312
TCARI	-0,882	0,400	0,416	-0,987	-0,971	0,846	0,854	0,895	0,889	0,808	0,842	-0,904	-0,908	1,000	1,000	0,781	0,774	-0,133	-0,268
MCARI	-0,882	0,400	0,416	-0,987	-0,971	0,846	0,854	0,895	0,889	0,808	0,842	-0,904	-0,908	1,000	1,000	0,781	0,774	-0,133	-0,268
CSI	-0,767	0,768	0,766	-0,793	-0,850	0,785	0,801	0,933	0,943	0,869	0,954	-0,848	-0,866	0,781	0,781	1,000	0,999	-0,323	-0,484
CSI_RE3	-0,759	0,774	0,765	-0,787	-0,845	0,777	0,794	0,928	0,941	0,862	0,951	-0,841	-0,862	0,774	0,774	0,999	1,000	-0,329	-0,487
B8_glcm	0,113	-0,409	-0,393	0,153	0,164	-0,108	-0,115	-0,250	-0,264	-0,146	-0,206	0,188	0,204	-0,133	-0,133	-0,323	-0,329	1,000	0,548
B11_glcm	0,251	-0,333	-0,318	0,295	0,304	-0,262	-0,272	-0,377	-0,390	-0,338	-0,416	0,295	0,312	-0,268	-0,268	-0,484	-0,487	0,548	1,000

Observando a Figura 33, o índice NBR2 aparece entre os seis primeiros atributos nas três métricas aplicadas e o índice CSI_RE3 entre os sete primeiros também nas três métricas. O índice NDWI_RE3 apareceu na quarta posição em ambas as métricas *ClassAE* e *CorrAE* e com valor também significativo em *GRAE*. Para a imagem de

abril/2019, a métrica *CorrAE* também apresentou as bandas RE3 e NIR e principalmente as imagens de textura, como sendo de baixa importância.

Semelhante aos valores da matriz de correlação de janeiro, a bandas SWIR1 e NIR de abril de 2019 resultaram em imagens de textura que tiveram as menores correlações em relação aos demais atributos, seguidas também das bandas RE3 e NIR. Os demais atributos apresentaram forte correlação entre si.

Finalmente, na Tabela 11 estão indicados os valores de importância dos atributos resultantes da aplicação das métricas de seleção, e na Figura 34, a matriz de correlação dos atributos referente à cena tomada no mês de julho/2018.

Tabela 11 - *Ranking* de Importância dos atributos definido pelas métricas *Classifier Attribute Evaluation (ClassAE)*, *Gain Ratio Attribute Evaluation (GRAE)* e *Correlation Attribute Evaluation (CorrAE)* para a imagem de julho/2018

ClassAE		GRAE		CorrAE	
CSI_RE3	0,4364	MSI	0,1945	NBR2	0,506
MSI_RE3	0,3805	CSI	0,1942	NDWI_RE3	0,448
NDWI_RE3	0,3664	NDWI	0,194	NDWI	0,4446
NBR2	0,3244	CSI_RE3	0,1854	MSI_RE3	0,4412
NDVI_RE3	0,2994	NDVI	0,182	MSI	0,4303
SWIR2	0,2955	NBR2	0,1813	CSI	0,4284
NDVI	0,2953	MSI_RE3	0,1736	TCARI	0,4251
NDWI	0,29	NDWI_RE3	0,1735	NDVI_RE3	0,4233
MSI	0,2839	NDVI_RE3	0,1703	CSI_RE3	0,4225
VERMELHO	0,2773	VERMELHO	0,1513	SWIR2	0,4155
CSI	0,273	NRI	0,1457	NDVI	0,4097
NRI	0,256	SWIR2	0,1313	NRI	0,4033
SWIR1	0,2306	MCARI	0,0873	VERMELHO	0,3664
RE3	0,2056	SWIR1	0,084	SWIR1	0,3648
MCARI	0,1915	NIR	0,0815	RE3	0,3581
TCARI	0,175	TCARI	0,0781	NIR	0,3425
NIR	0,1592	RE3	0,0689	MCARI	0,3251
SWIR1_glcml	0,046	NIR_glcml	0,024	SWIR1_glcml	0,223
NIR_glcml	-0,0235	SWIR1_glcml	0,0208	NIR_glcml	0,0535

Figura 34 - Matriz de correlação dos atributos referente ao mês de julho/2018

	Red	RE3	IVP	SWIR1	SWIR2	NDVI	NDVI_RE3	NDWI	NDWI_RE3	NRI	NBR2	MSI	MSI_RE3	TCARI	MCARI	CSI	CSI_RE3	IVP_glc	SWIR1_glc
Red	1,000	-0,577	-0,611	0,848	0,938	-0,984	-0,983	-0,907	-0,889	-0,947	-0,879	0,942	0,927	-0,882	-0,882	-0,767	-0,759	0,113	0,251
RE3	-0,577	1,000	0,990	-0,380	-0,542	0,649	0,660	0,738	0,749	0,696	0,707	-0,688	-0,702	0,400	0,400	0,768	0,774	-0,409	-0,333
IVP	-0,611	0,990	1,000	-0,391	-0,557	0,683	0,689	0,753	0,753	0,722	0,713	-0,710	-0,714	0,416	0,416	0,766	0,765	-0,393	-0,318
SWIR1	0,848	-0,380	-0,391	1,000	0,965	-0,806	-0,817	-0,887	-0,885	-0,776	-0,835	0,883	0,892	-0,987	-0,987	-0,793	-0,787	0,153	0,295
SWIR2	0,938	-0,542	-0,557	0,965	1,000	-0,919	-0,927	-0,952	-0,946	-0,891	-0,924	0,958	0,960	-0,971	-0,971	-0,850	-0,845	0,164	0,304
NDVI	-0,984	0,649	0,683	-0,806	-0,919	1,000	0,999	0,926	0,907	0,961	0,896	-0,959	-0,944	0,846	0,846	0,785	0,777	-0,108	-0,262
NDVI_RE3	-0,983	0,660	0,689	-0,817	-0,927	0,999	1,000	0,936	0,920	0,965	0,907	-0,964	-0,952	0,854	0,854	0,801	0,794	-0,115	-0,272
NDWI	-0,907	0,738	0,753	-0,887	-0,952	0,926	0,936	1,000	0,997	0,923	0,953	-0,978	-0,982	0,895	0,895	0,933	0,928	-0,250	-0,377
NDWI_RE3	-0,889	0,749	0,753	-0,885	-0,946	0,907	0,920	0,997	1,000	0,911	0,954	-0,967	-0,979	0,889	0,889	0,943	0,941	-0,264	-0,390
NRI	-0,947	0,696	0,722	-0,776	-0,891	0,961	0,965	0,923	0,911	1,000	0,945	-0,907	-0,899	0,808	0,808	0,869	0,862	-0,146	-0,338
NBR2	-0,879	0,707	0,713	-0,835	-0,924	0,896	0,907	0,953	0,954	0,945	1,000	-0,907	-0,915	0,842	0,842	0,954	0,951	-0,206	-0,416
MSI	0,942	-0,688	-0,710	0,883	0,958	-0,959	-0,964	-0,978	-0,967	-0,907	-0,907	1,000	0,996	-0,904	-0,904	-0,848	-0,841	0,188	0,295
MSI_RE3	0,927	-0,702	-0,714	0,892	0,960	-0,944	-0,952	-0,982	-0,979	-0,899	-0,915	0,996	1,000	-0,908	-0,908	-0,866	-0,862	0,204	0,312
TCARI	-0,882	0,400	0,416	-0,987	-0,971	0,846	0,854	0,895	0,889	0,808	0,842	-0,904	-0,908	1,000	1,000	0,781	0,774	-0,133	-0,268
MCARI	-0,882	0,400	0,416	-0,987	-0,971	0,846	0,854	0,895	0,889	0,808	0,842	-0,904	-0,908	1,000	1,000	0,781	0,774	-0,133	-0,268
CSI	-0,767	0,768	0,766	-0,793	-0,850	0,785	0,801	0,933	0,943	0,869	0,954	-0,848	-0,866	0,781	0,781	1,000	0,999	-0,323	-0,484
CSI_RE3	-0,759	0,774	0,765	-0,787	-0,845	0,777	0,794	0,928	0,941	0,862	0,951	-0,841	-0,862	0,774	0,774	0,999	1,000	-0,329	-0,487
B8_glc	0,113	-0,409	-0,393	0,153	0,164	-0,108	-0,115	-0,250	-0,264	-0,146	-0,206	0,188	0,204	-0,133	-0,133	-0,323	-0,329	1,000	0,548
B11_glc	0,251	-0,333	-0,318	0,295	0,304	-0,262	-0,272	-0,377	-0,390	-0,338	-0,416	0,295	0,312	-0,268	-0,268	-0,484	-0,487	0,548	1,000

Observando a Tabela 11 verifica-se que, para as três métricas, somente o índice de vegetação NBR2 aparece entre os seis atributos de maior importância. Os índices MSI_RE3, NDWI_RE3, NBR2 e NDVI_RE3 também obtiveram posições significativas para as três métricas aplicadas. Para a imagem de julho/2018, as imagens de textura e as bandas do NIR e RE3, também foram apresentadas como de baixa importância nas três métricas.

A Figura 34 também mostra as imagens de textura com baixas correlações em relação aos demais atributos, seguidas das bandas RE3 e NIR. Os demais atributos apresentaram forte correlação entre eles. O índice NRI apresentou correlações relativamente baixas em relação aos atributos RE3, NIR, SWIR1, TCARI e MCARI. Em relação aos demais atributos, o índice NRI apresentou correlações altas com valores a partir de 0,774 com o índice CSI_RE3.

De maneira geral, o índice CSI_RE3 foi único que apareceu nas três métricas de seleção de atributos no primeiro conjunto de amostras e os índices CSI_RE3, NBR2 e NDWI_RE3 foram os atributos de maior importância definidos pelas métricas aplicadas ao segundo conjunto de amostras.

As matrizes de correlação referentes aos atributos coletados do segundo conjunto de amostras apresentaram resultados semelhantes para as três datas, com baixos valores de correlação para os atributos de textura em relação aos demais, seguido das bandas RE3 e NIR. No geral, os índices utilizados apresentaram altas correlação entre si.

A métrica *CorrAE* apresentou bastante coerência entre o seu resultado e a matriz de correlação criada para o primeiro conjunto de amostras. Para o segundo conjunto de amostras, os atributos de textura, apesar de apresentarem baixos valores de

intercorrelações, ocuparam as últimas posições nos “*rankings* de importância” gerados pela métrica, indicando que pouco contribuem na discriminação entre as classes.

6.1.3 Definição dos conjuntos de atributos

Os experimentos 1 e 2 auxiliaram na definição de cinco conjuntos de atributos, para uma posterior aplicação do conjunto de melhor desempenho no processo de classificação. Os índices de vegetação calculados com a substituição da banda do NIR pela do RE3 foram utilizados, desconsiderando-se as formulações originais.

Os conjuntos de atributos definidos para desenvolvimento das classificações pelos algoritmos RNA e FA foram:

- Conjunto 1: todos os atributos foram definidos como dados de entrada;
- Conjunto 2: foram escolhidos aqueles atributos que obtiveram os melhores resultados nas métricas de seleção de atributos aplicadas, tanto no experimento 1, quanto no 2;
- Conjunto 3: inclui os atributos que definiram as maiores diferenças nas médias das amostras coletadas na imagem de julho de 2018, referente ao experimento 1;
- Conjunto 4: refere-se às bandas espectrais e imagens de textura e aos índices espectrais normalizados;
- Conjunto 5: semelhante ao conjunto 4, porém utilizando os índices não-normalizados.

A Tabela 12 resume a composição dos cinco conjuntos de atributos definidos:

Tabela 12 – Composição dos conjuntos de atributos definidos para classificação

Conjunto	Atributos de entrada
1	<u>Bandas espectrais:</u> Red, RE3, NIR, SWIR1 e SWIR2
	<u>Imagens de textura:</u> NIR_glcm e SWIR1_glcm
	<u>Índices espectrais:</u> NDVI_RE3, NDWI_RE3, NRI, NBR2, MSI_RE3, MCARI, TCARI e CSI_RE3
2	<u>Bandas espectrais:</u> RE3, NIR e SWIR1
	<u>Índices espectrais:</u> NDWI_RE3, NRI, NBR2 e CSI_RE3
3	<u>Bandas espectrais:</u> RE3 e SWIR1
	<u>Imagens de textura:</u> SWIR1_glcm
	<u>Índices espectrais:</u> NDVI_RE3, NDWI_RE3 e CSI_RE3
4	<u>Bandas espectrais:</u> Red, RE3, NIR, SWIR1 e SWIR2
	<u>Imagens de textura:</u> NIR_glcm e SWIR1_glcm
	<u>Índices espectrais:</u> NDVI_RE3, NDWI_RE3, NRI e NBR2
5	<u>Bandas espectrais:</u> Red, RE3, NIR, SWIR1 e SWIR2
	<u>Imagens de textura:</u> NIR_glcm e SWIR1_glcm
	<u>Índices espectrais:</u> MSI_RE3, MCARI, TCARI e CSI_RE3

Esses conjuntos de atributos foram utilizados para o treinamento dos algoritmos RNA e FA, avaliando-se o desempenho de cada um em cada fase fenológica do café (julho de 2018, janeiro e abril de 2019) e aplicando-os no processo de classificação.

6.2 Resultado da aplicação dos algoritmos de classificação

Para o treinamento dos algoritmos FA e RNA foram utilizadas as amostras adquiridas no experimento 2. Lembrando que essas amostras foram coletadas com base no segundo conjunto de dados de referência disponibilizadas pela UFU e nas áreas de café mapeadas pela Emater-MG.

Para cada conjunto de atributos definido na Tabela 12, foram gerados três arquivos ARFF, um para cada data (julho/2018, janeiro e abril/2019), constituindo um total de 15 arquivos ARFF's. Cada arquivo possui somente os valores dos pixels amostrados para um específico conjunto de atributos de cada uma das datas de tomada da imagem.

Para o processo de treinamento foram utilizados os arquivos ARFF's criados e especificados os parâmetros dos algoritmos de classificação. A RNA foi configurada com duas camadas escondidas, sendo que o número de elementos de processamento (nós) da primeira camada escondida foi estabelecido pela soma do número de atributos de entrada e classes de saída, e da segunda camada escondida pela metade desse valor. Os parâmetros de treinamento adotados foram: taxa de aprendizado (α) igual 0,01, termo de momento igual 0,1, função de ativação sigmodal, e número de ciclos de treinamento igual 3000.

Para o treinamento de FA, o número de árvores (*ntree*) ficou estabelecido em 1000 e o número de variáveis amostradas aleatoriamente como candidatas em cada divisão (*mtry*) foi definido pela raiz quadrada do número de camadas disponíveis. A Tabela 13 mostra os números de nós para as duas camadas escondidas de cada conjunto de atributos para o treinamento da RNA, assim como os valores do parâmetro *mtry* definidos para o algoritmo FA, também para cada conjunto.

Tabela 13 – Número de nós das camadas escondidas da RNA e valores do parâmetro *mtry* do FA, referentes aos cinco conjuntos de atributos

Conjunto	Camadas escondidas	<i>mtry</i>
	RNA	RF
1	18; 9	3
2	10; 5	2
3	9; 5	2
4	14; 7	3
5	14; 7	3

O desempenho dos algoritmos sobre o conjunto de dados de validação, para cada conjunto de atributos e cada classificador, foi comparado a partir dos valores de *Root Mean Square Error* (RMSE), disponibilizados no produto de saída de cada processamento. A partir da análise desse valor, foi selecionado o conjunto de atributo mais adequado para a aplicação do algoritmo de classificação. Vale ressaltar que os treinamentos foram realizados com 80% dos pixels amostrados, de modo que o conjunto de validação considerou 20% da amostra original, os quais foram definidos de forma randômica pelo WEKA. A Tabela 14 mostra os valores de RMSE obtidos a partir da classificação dos dados de validação pelos algoritmos FA e RNA, para os diferentes conjuntos de atributos de entrada e datas de aquisição das imagens.

Tabela 14 – Valores de RMSE obtidos a partir das classificações dos dados de validação (20% da amostra) para os diferentes conjuntos de atributos e datas de aquisição das imagens

Conjunto	Data	FA	RNA
		RMSE - Validação	RMSE - Validação
1	jan/19	0,1042	0,1398
	abr/19	0,0915	0,1294
	jul/18	0,0727	0,0851
2	jan/19	0,1179	0,1852
	abr/19	0,1131	0,1978
	jul/18	0,0978	0,1666
3	jan/19	0,1061	0,1862
	abr/19	0,1080	0,1884
	jul/18	0,0935	0,1560
4	jan/19	0,1147	0,1669
	abr/19	0,1156	0,1614
	jul/18	0,0898	0,1267
5	jan/19	0,0983	0,1391
	abr/19	0,0852	0,1404
	jul/18	0,0682	0,0906

Pode-se verificar pela Tabela 14 que, em todos os casos, os valores de RMSE obtidos foram menores para o classificador FA. Também se observa que, para todos os conjuntos de dados de entrada classificados pelos dois algoritmos, a imagem referente a julho de 2018 resultou em menores valores de RMSE.

De acordo com os resultados referentes às imagens de julho dos 5 conjuntos, os menores valores de RMSE foram obtidos para aqueles conjuntos com maiores quantidades de atributos de entrada (conjuntos 1 e 5 com 15 e 11 atributos, respectivamente). Os menores valores entre eles, foi obtido pelo conjunto 5 para o FA e pelos conjuntos 1 e 5 para RNA. Apesar desses conjuntos terem apresentado os menores valores de RMSE, observa-se que, os conjuntos com menores quantidades de atributos obtiveram resultados bastante próximos, o que pode ser vantajoso utilizá-los em termos de tempo de processamento.

Os resultados mostraram que todos os conjuntos de atributos se adequaram melhor à fase fenológica referente ao mês de julho. Além disso, o conjunto 5 foi o mais adequado para a aplicação da classificação nas três fases fenológicas do café tratadas. Assim, foi definido o conjunto 5 para a aplicação no processo de classificação. Esse conjunto foi utilizado para a classificação das imagens de julho/2018 e janeiro e abril/2019 utilizando os dois classificadores. Para esse processo foi utilizado o modelo salvo após o treinamento das datas do conjunto.

6.2.1 Desempenho dos algoritmos utilizados na classificação das imagens multitemporais

A Tabela 15 mostra algumas estatísticas geradas dos resultados obtidos nos processos de treinamento dos classificadores FA e RNA utilizando o conjunto 5, para uma melhor comparação dos seus desempenhos. Essas estatísticas são o índice *Kappa* e a exatidão global (EG). Em seguida, a Tabela 16 mostra os valores de *F-Score* gerados nos mesmos processamentos.

Tabela 15 – Indicadores estatísticos da validação e tempo de processamento do treinamento dos classificadores FA e RNA referentes ao conjunto de atributos 5

	FA			RNA		
	<i>Kappa</i>	EG	Tempo de processamento	<i>Kappa</i>	EG	Tempo de processamento
jan/19	0,9779	0,9882	01 min 49,12 segs.	0,9300	0,9627	12 min 11,15 segs.
abr/19	0,9824	0,9906	01 min 40,02 segs.	0,9129	0,9541	12 min 29,36 segs.
jul/18	0,9889	0,9941	01 min 43,71 segs.	0,9710	0,9846	12 min 28,94 segs.

Tabela 16 – Valores de *F-Score* gerados no treinamento dos classificadores FA e RNA referentes ao conjunto de atributos 5

		<i>F-Score</i>		
		Café infectado	Café sadio	Solo exposto
FA	jan/2019	0,987	0,989	1,000
	abr/2019	0,990	0,991	0,993
	jul/2018	0,993	0,994	1,000
RNA	jan/2019	0,959	0,964	1,000
	abr/2019	0,950	0,965	0,839
	jul/2018	0,983	0,985	1,000

Os valores encontrados na Tabela 15 confirmam que o classificador FA obteve melhor desempenho como mostra os indicadores estatísticos índice *Kappa*, EG, e o tempo de processamento para o treinamento das amostras, que foi consideravelmente menor. Além disso, os treinamentos do classificador FA proporcionaram resultados satisfatórios, uma vez que apresentaram valores de índice *Kappa* superiores a 0,97. Destaca-se também, os valores mais baixos obtidos para os índices *Kappa* referentes às cenas de abril (0,9129) e janeiro (0,9300), ainda que satisfatórios.

Os valores referentes à medida *F-Score* (Tabela 16) indicam a qualidade geral dos modelos aplicados. De modo geral, o classificador FA obteve resultados satisfatórios para as três classes, inclusive atingindo valor máximo para a classe de solo exposto nos meses

de janeiro/2019 e julho/2018. Os demais valores obtidos pelo classificador foram superiores a 0,98 para as três classes.

Para o classificador RNA, destaca-se o valor mais baixo obtido de 0,839 para a classe de solo exposto do mês de abril. Os demais valores de *F-Score* para esse classificador ficaram acima de 0,95.

O próximo passo foi a realização da classificação das imagens de cada data de tomada utilizando o conjunto de atributos 5 e os dois classificadores. A Figura 35 e a Figura 36 apresentam as matrizes de confusão geradas após a classificação referentes aos algoritmos FA e RNA, respectivamente, em que AU se refere à acurácia do usuário e AP à acurácia do produtor. Lembrando que a matriz de confusão foi computada a partir dos 20% dos dados de validação da amostra, sendo definidos de forma aleatória pelo WEKA.

Figura 35 - Matrizes de confusão geradas a partir dos dados de validação do classificador FA das cenas de (a) janeiro/2019, (b) abril/2019 e (c) julho/2018, em que CI = café infectado, CS = café sadio e SE = solo exposto

(a)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2799	21	0	2820	0,993
	CS	53	3217	0	3270	0,984
	SE	0	0	223	223	1,000
	Total	2852	3238	223	6313	
AP		0,981	0,994	1,000		

(b)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2829	36	0	2865	0,987
	CS	20	3202	0	3222	0,994
	SE	3	0	223	226	0,987
	Total	2852	3238	223	6313	
AP		0,992	0,989	1,000		

(c)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2762	16	0	2778	0,994
	CS	21	3289	0	3310	0,994
	SE	0	0	225	225	1,000
	Total	2783	3305	225	6313	
AP		0,992	0,995	1,000		

Figura 36 - Matrizes de confusão geradas a partir dos dados de validação do classificador RNA Artificiais das cenas de (a) janeiro/2019, (b) abril/2019 e (c) julho/2018, em que CI = café infectado, CS = café sadio e SE = solo exposto.

(a)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2742	125	0	2867	0,956
	CS	110	3113	0	3223	0,966
	SE	0	0	223	223	1,000
	Total	2852	3238	223	6313	
AP		0,961	0,961	1,000		

(b)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2748	124	62	2934	0,937
	CS	104	3114	0	3218	0,968
	SE	0	0	161	161	1,000
	Total	2852	3238	223	6313	
AP		0,964	0,962	0,722		

(c)		Referência			Total	AU
		CI	CS	SE		
Classificação	CI	2745	59	0	2804	0,979
	CS	38	3246	0	3284	0,988
	SE	0	0	225	225	1,000
	Total	2783	3305	225	6313	
AP		0,986	0,982	1,000		

De acordo com a Figura 35, as classes de café infectado e sadio mostram uma pequena confusão, o que pode ser atribuído ao comportamento espectral bastante similar entre essas ocorrências. Porém, de maneira geral, os valores das acurácias do usuário (AU) e do produtor (AP) obtidos para o FA, para todas as classes, foram superiores a 0,98, para as três fases fenológicas.

A classe de solo exposto foi a melhor identificada nas cenas de janeiro/2019 e julho/2018, com valores das acurácias iguais a 1 tanto para a acurácia do produtor quanto do usuário, demonstrando que não houve confusão entre áreas de solo exposto e talhões de café infectados, que algumas vezes, possuem algumas falhas entre as plantas apresentando semelhança espectral com áreas de solo exposto.

Em relação aos resultados obtidos com o classificador RNA (Figura 36), os valores de AU e AP permaneceram acima de 0,93, com exceção da AP obtida em relação à classe de solo exposto da cena de abril/2019, em que boa parte dos pixels foram associados à classe café infectado. Esse resultado é consistente com o baixo valor de F -

Score obtido para a classe (0,839) (Tabela 16). A presença de reboleiras ou clareiras causadas pelos nematoides em talhões de café infectado, podem ser a causa dessa confusão. Adicionalmente, o conjunto escolhido, pode não ter se adequado muito bem às fases fenológicas referentes ao mês de abril.

Os resultados obtidos para o classificador FA foram semelhantes para as imagens das três fases fenológicas do café. Já nos obtidos pela RNA, a cena de abril/2019 foi a que resultou em menores acurácias, ainda que satisfatórias, enquanto que a imagem de julho/2018, no melhor desempenho.

6.2.2 Mapas temáticos resultantes das classificações multitemporais

A distribuição espacial das classes obtida a partir da aplicação dos algoritmos FA e RNA ao conjunto de atributos de melhor desempenho na classificação das cenas de julho/2018 e janeiro e abril/2019 é apresentada na forma de mapas temáticos elaborados para cada uma das datas de tomada da imagem. Além disso, para cada cenário analisado, foi produzida uma imagem de incerteza a partir do resultado da classificação obtido para cada pixel. A informação sobre incerteza permite identificar as áreas com maior probabilidade de erro no processo de tomada de decisão do classificador, sendo essa informação útil para representar e identificar os padrões espaciais dos erros da classificação.

Na sequência são apresentados os mapas temáticos resultantes da classificação por FA. A Figura 37 mostra o mapa temático obtido para a cena de janeiro/2019, seguido da Figura 38, para a cena de abril/2019 e a Figura 39, para a cena de julho/2018. As imagens de incerteza são mostradas nas Figura 40, 41 e 42 para as cenas de janeiro/2019, abril/2019 e julho/2018, respectivamente.

Figura 37 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de janeiro/2019

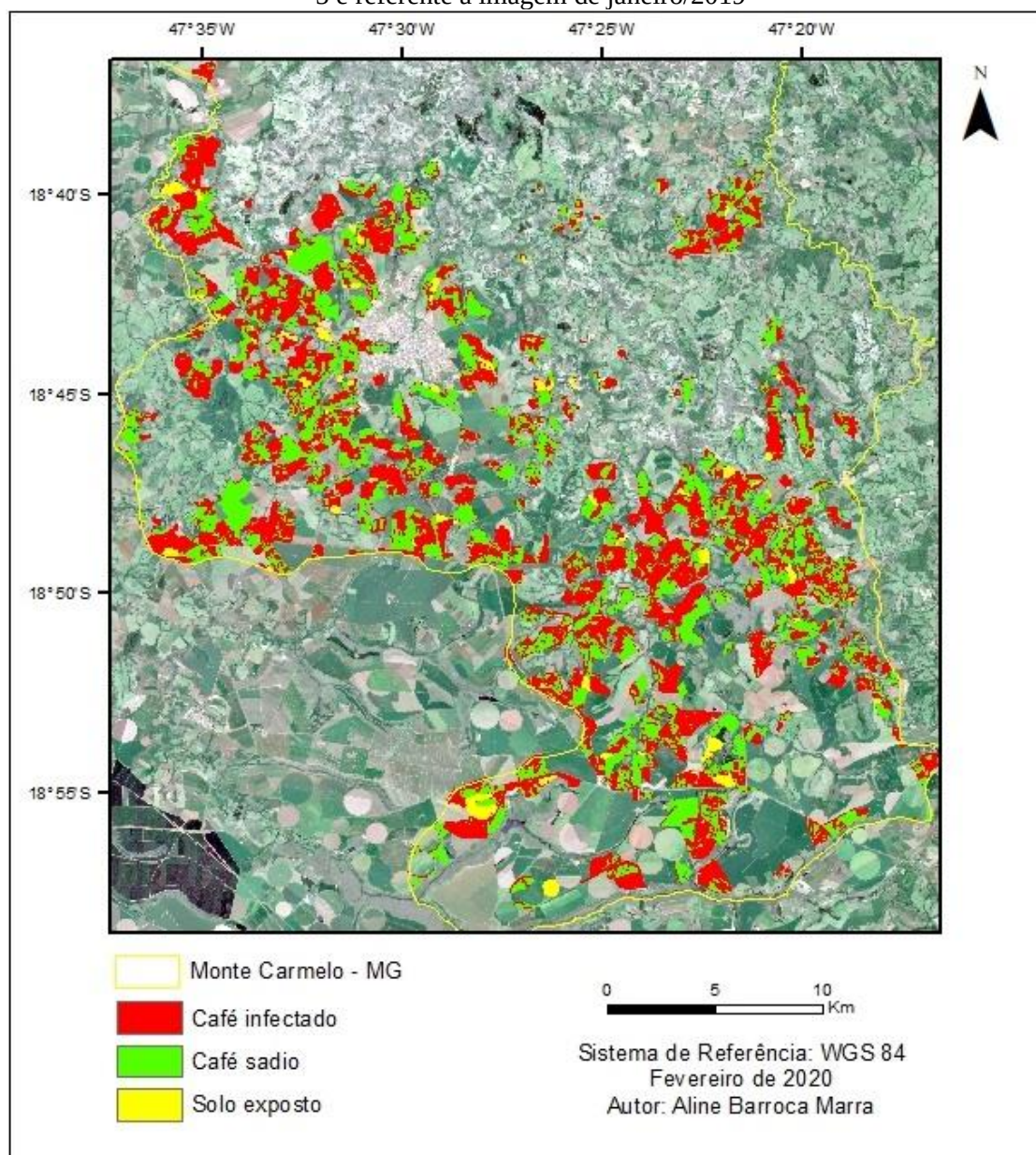


Figura 38 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de abril/2019

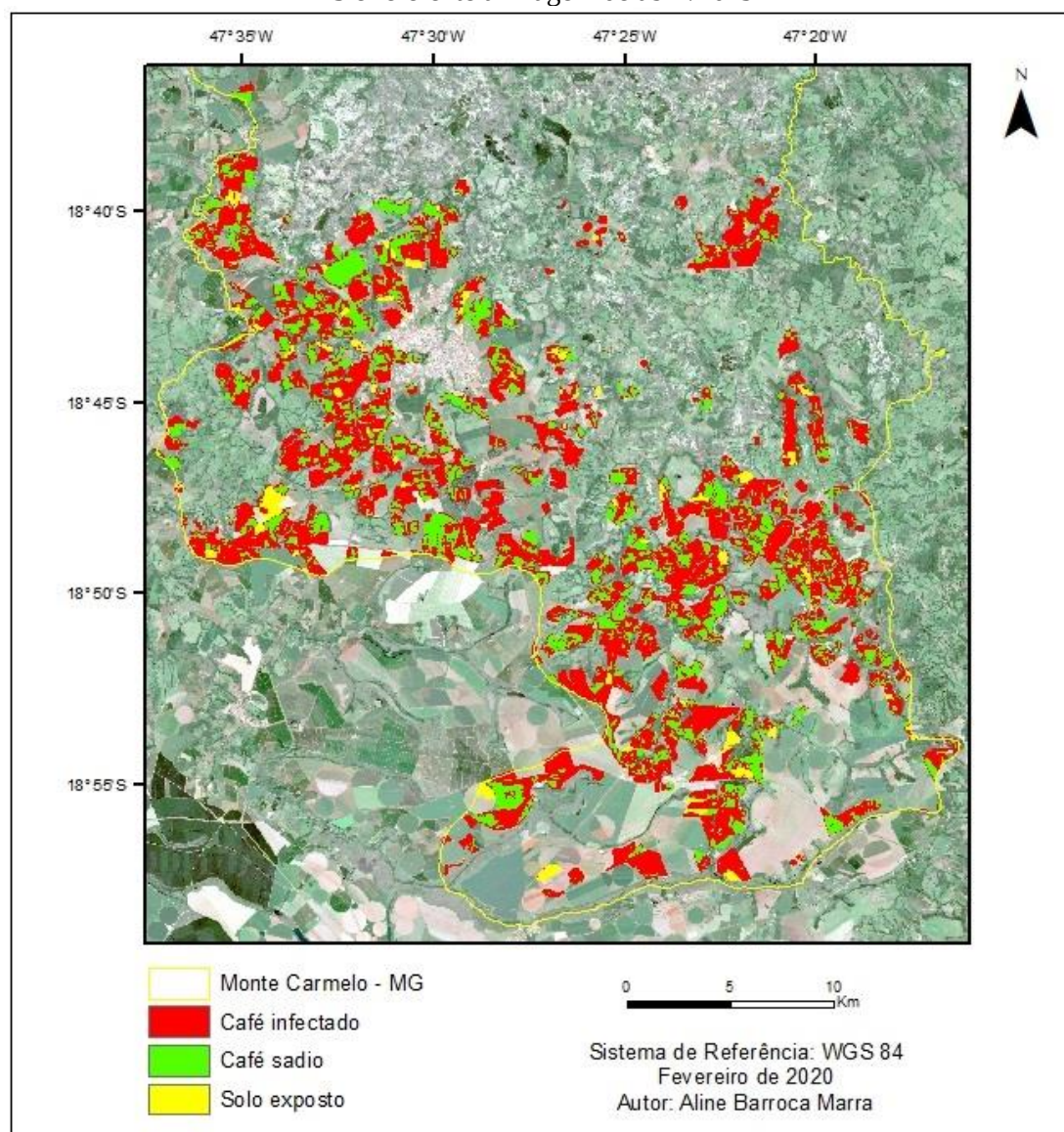


Figura 39 – Mapa temático resultante da classificação por FA, aplicado ao conjunto de atributos 5 e referente à imagem de julho/2018

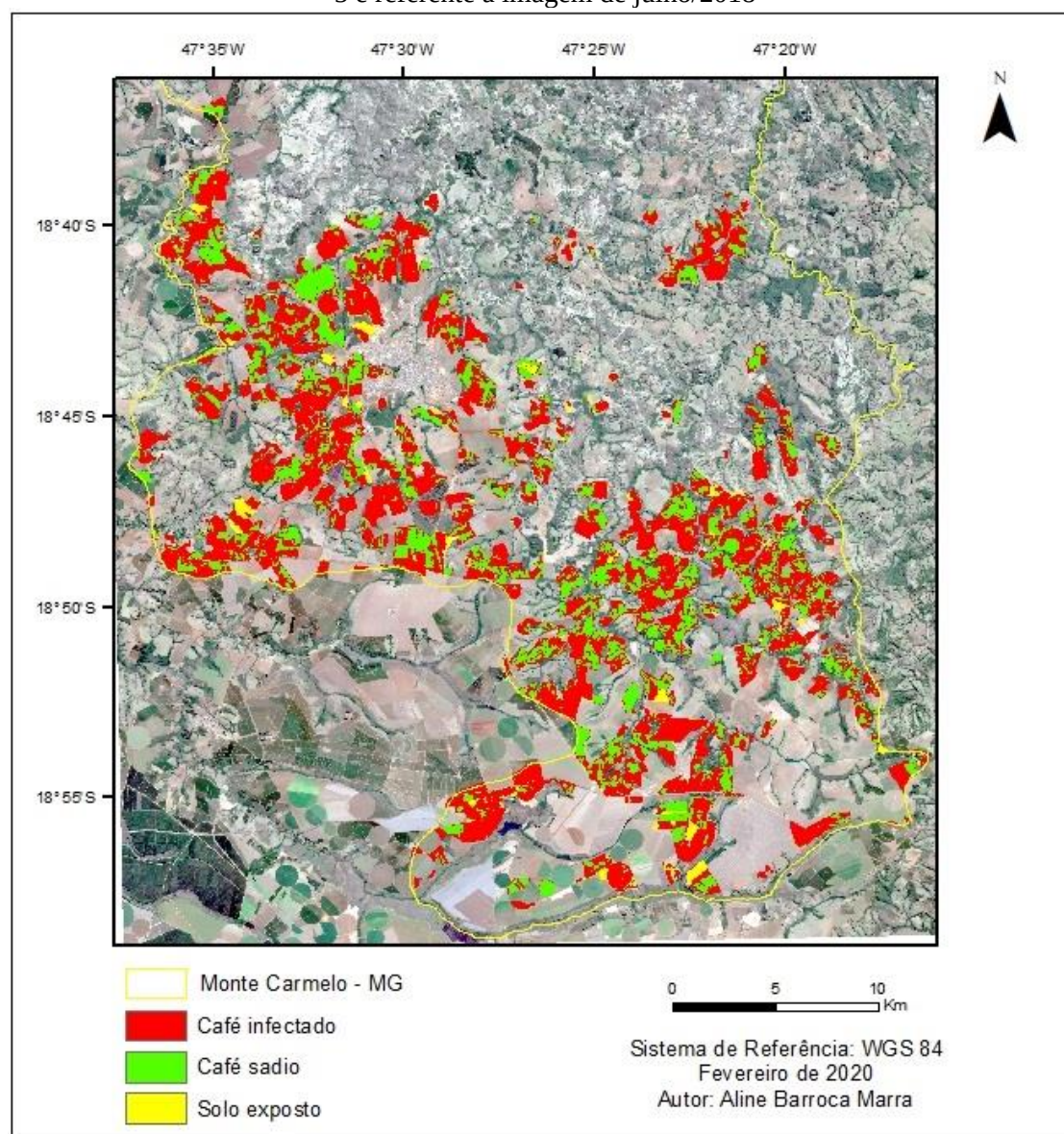


Figura 40 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a janeiro/2019

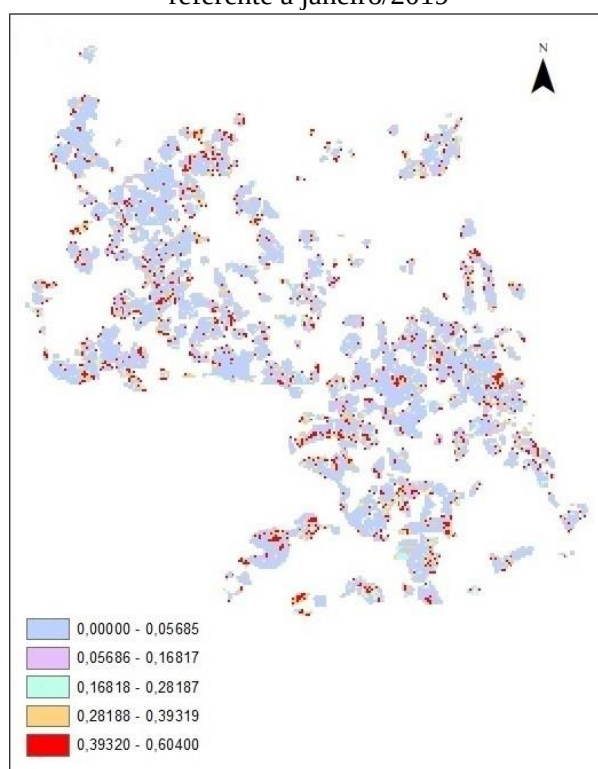


Figura 41 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a abril/2019

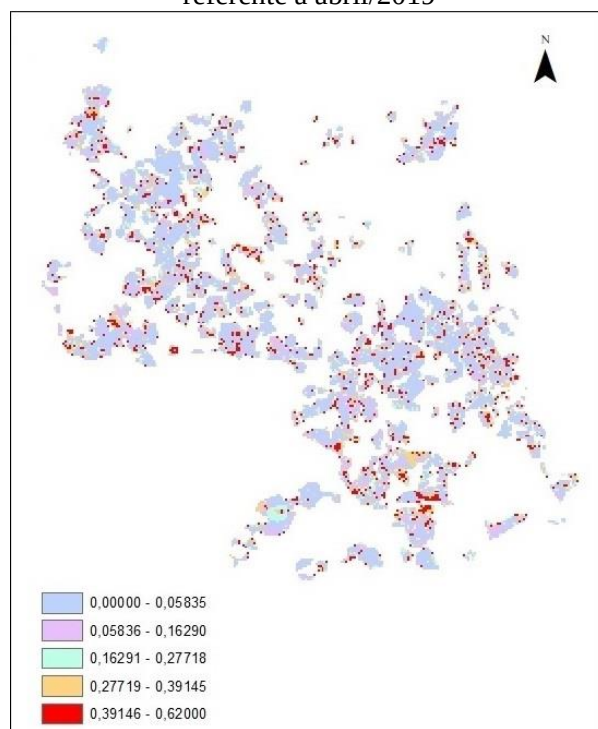
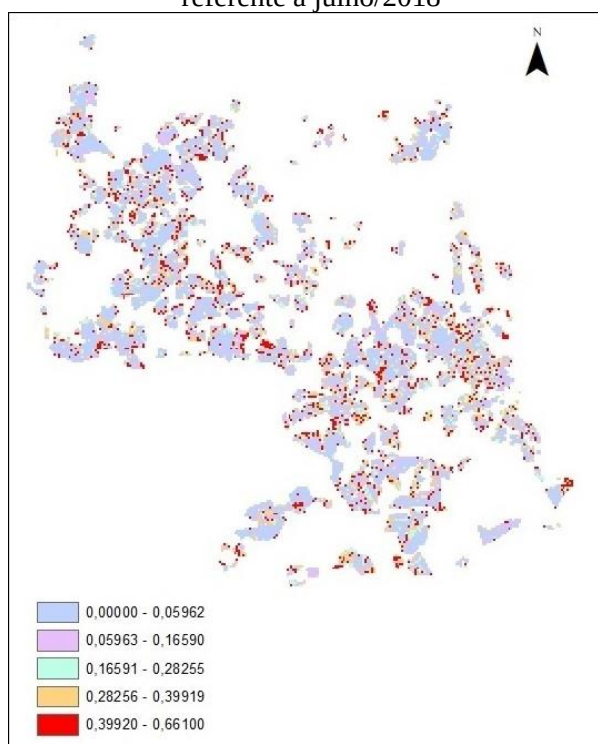


Figura 42 – Imagem de incerteza gerada a partir do resultado da classificação utilizando o FA, referente a julho/2018



Em uma primeira análise é possível verificar que todas as classes foram representadas nos mapas mostrados nas Figuras 37, 38 e 39, porém, verifica-se que a classe café infectado foi predominante nas classificações das três cenas.

A imagem de janeiro/2019 apresenta menor predominância de café infectado em relação as outras datas. A alta precipitação e temperatura moderada são características dessa época, o que oferece maior vigor à planta. Além disso, a ocorrência de reboleiras causadas pelos nematoides são mais comuns em talhões de café irrigados. Por se tratar de uma área de estudo com predominância de cultivo de café irrigado, a diferença espectral se torna mais sutil entre áreas de café infectado por nematoides e áreas de café sadio, podendo assim, mascarar a ocorrência de nematoides.

Por outro lado, em condições de seca e baixas temperaturas, características de julho, por se tratar das fases fenológicas de repouso e autopoda, áreas sadias podem ser rotuladas como áreas infectadas, ou então, podem oferecer uma condição mais livre da influência de muita chuva.

As imagens de incerteza, Figura 40, 41 e 42, geradas a partir das classificações das cenas de janeiro/2019, abril/2019 e julho/2018, respectivamente, apresentaram distribuições semelhantes entre as datas para o classificador FA, em que as áreas em

vermelho são as localizações das maiores incertezas da classificação. As imagens também mostraram a predominância de baixa incerteza (tom azul nas imagens).

Os mapas temáticos gerados a partir dos resultados da aplicação da RNA, utilizando o conjunto de atributo mais adequado para classificação das cenas de janeiro/2019, abril/2019 e julho/2018, são apresentados nas Figuras 43, 44, e 45, respectivamente. Já as imagens de incerteza, podem ser vistas nas Figuras 46, 47 e 48, referentes a janeiro/2019, abril/2019 e julho/2018, respectivamente.

Figura 43 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem janeiro/2019

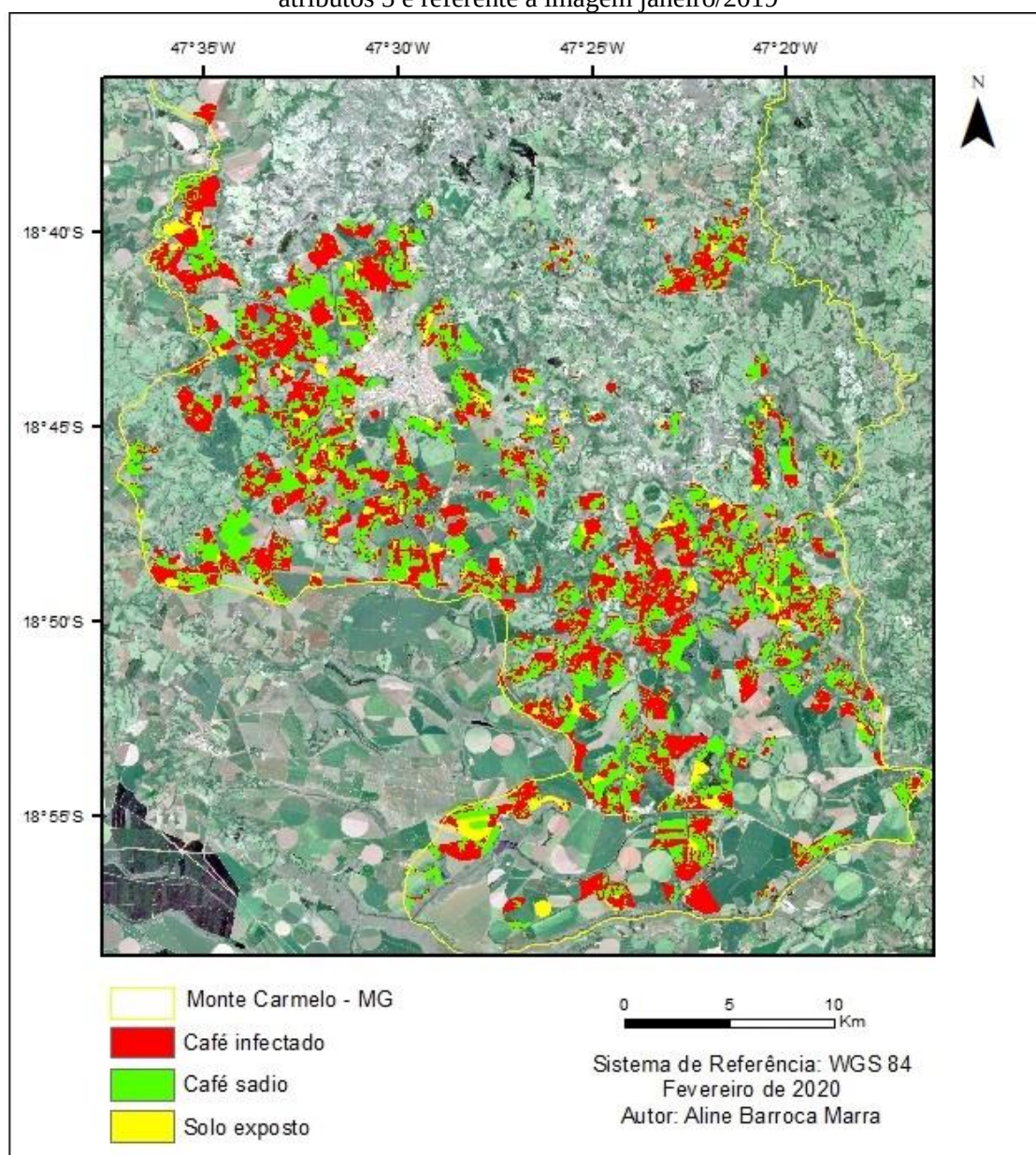


Figura 44 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem abril/2019

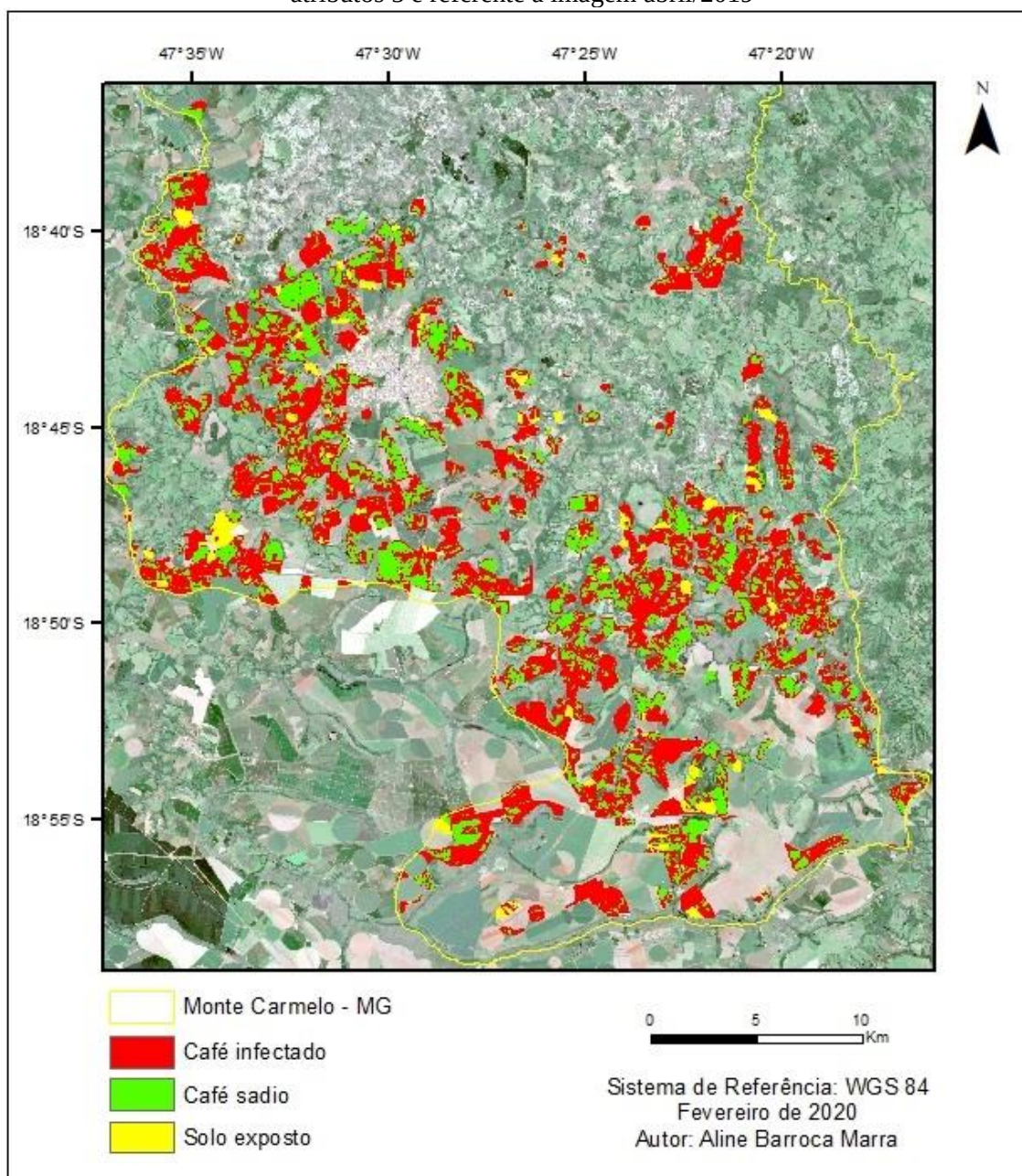


Figura 45 – Mapa temático resultante da classificação por RNA, aplicado ao conjunto de atributos 5 e referente à imagem julho/2018

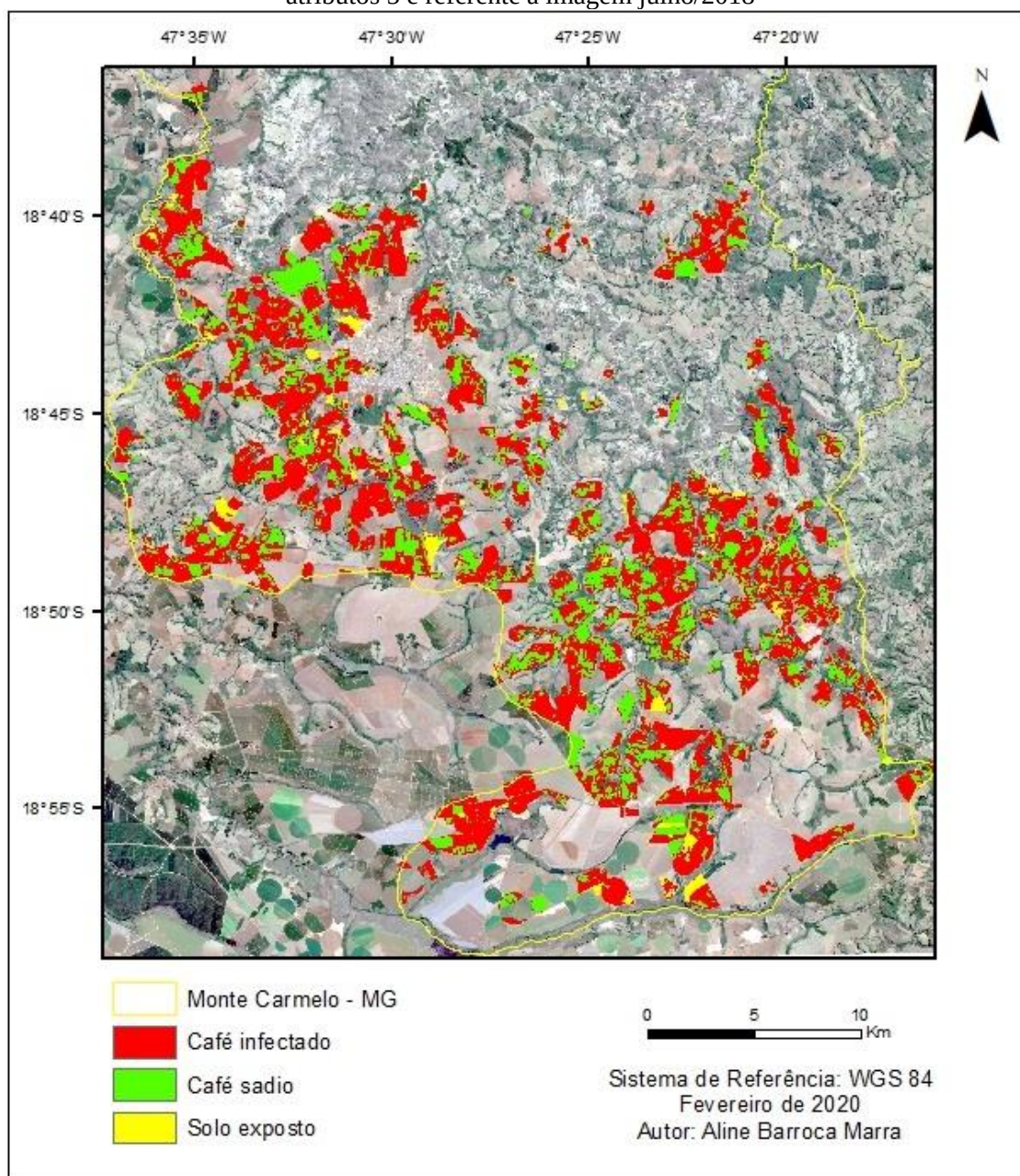


Figura 46 - Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a janeiro/2019

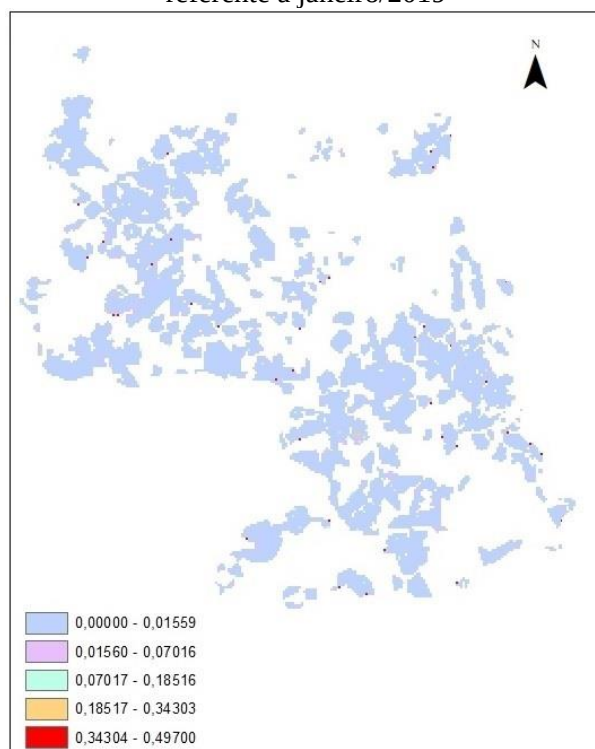


Figura 47 – Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a abril/2019

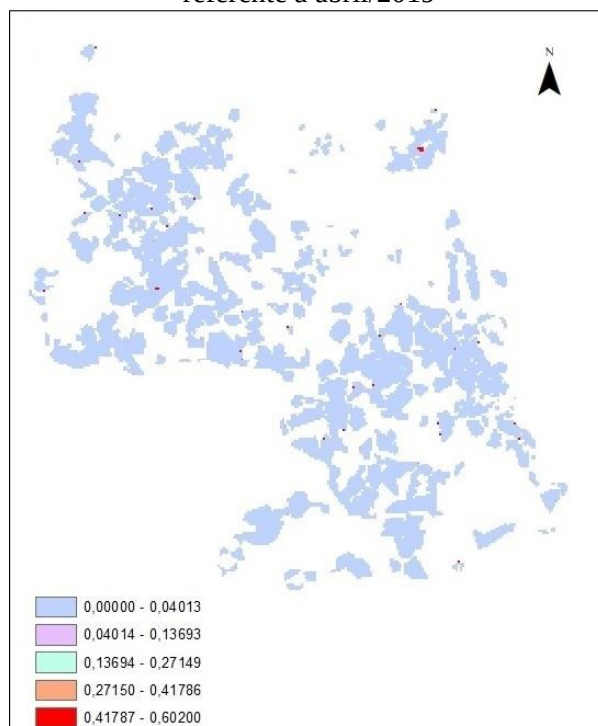
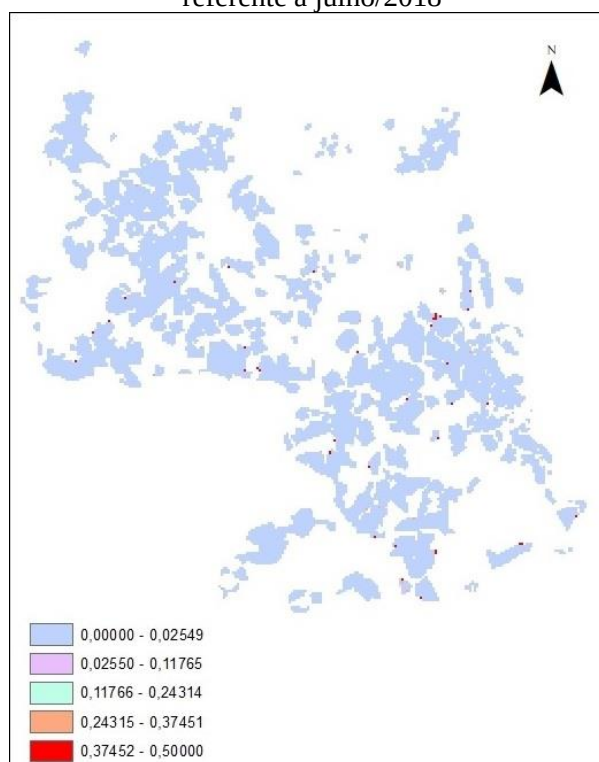


Figura 48 – Imagem de incerteza gerada a partir do resultado da classificação utilizando a RNA, referente a julho/2018



Os mapas gerados da classificação dos atributos utilizando o algoritmo RNA apresentaram similaridade com aqueles gerados pelo FA, com a predominância da classe de café infectado nas três datas. Os atributos derivados da cena de janeiro, também para RNA, definiram uma menor representatividade dessa classe, em relação à demais épocas avaliadas.

Apesar dos dois algoritmos resultarem em classificações visualmente similares, as imagens de incerteza geradas a partir da aplicação da RNA (Figuras 46, 47 e 48) mostram predominância de valores de baixas incertezas, apontando para um resultado mais provável quanto à atribuição de um pixel a uma dada classe.

Para uma melhor comparação entre os classificadores, foi calculada a proporção de pixels com maior incerteza para os diferentes métodos de classificação e datas de tomada das imagens (Tabela 17).

Tabela 17 – Proporção de pixels com maior incerteza na atribuição da classe para os diferentes métodos de classificação e datas de tomada das imagens

Data de aquisição da imagem	Método de classificação	
	FA	RNA
jan/19	Incerteza > 0,39320 12,95%	Incerteza > 0,34304 4,79%
abr/19	Incerteza > 0,39146 12,70%	Incerteza > 0,41787 2,38%
jul/18	Incerteza > 0,39920 13,18%	Incerteza > 0,37452 2,95%

Com base na incerteza na atribuição de cada classe, os valores apresentados na Tabela 17, indicam que, apesar do classificador FA apresentar melhor desempenho em relação aos resultados obtidos pelos indicadores estatísticos gerados e ao tempo de processamento, a RNA aponta para uma maior confiança no resultado da classificação para todas as datas de tomada de imagem, pois obteve proporções significativamente menores na atribuição das classes em relação ao FA. Enquanto RNA definiu proporções de pixels menores que 5% para as três datas de aquisição da imagem, as classificações resultantes da aplicação de FA resultou em proporções superiores a 12% de pixels com alta incerteza na sua atribuição a uma dada classe (taxa superior a 0,39).

Também para fins comparativos, foram calculadas as proporções de pixels associados a cada classe, para cada época de tomada da imagem considerando os dois métodos de classificação utilizados. As proporções obtidas são mostradas na Tabela 18.

Tabela 18 – Proporção de pixels classificados para cada classe para os diferentes métodos de classificação e datas de tomada das imagens

		Café infectado	Café sadio	Solo exposto
FA	jan/19	53,82%	41,93%	4,25%
	abr/19	61,34%	32,61%	6,05%
	jul/18	63,37%	32,40%	4,23%
RNA	jan/19	47,81%	44,31%	7,87%
	abr/19	59,41%	33,96%	6,62%
	jul/18	63,94%	31,54%	4,52%

Os valores apresentados na Tabela 18 confirmam a predominância de áreas classificadas como café infectado para ambos os métodos de classificação, nas três datas de tomada das imagens. Os resultados mostram que embora a área de estudos seja caracterizada pela predominância de café infectado, essa alta proporção de áreas

infectadas deve considerar que outras condições de estresse da planta podem definir um padrão espectral similar daquele característico da presença de nematoides.

Portanto, ainda que a classe café infectado possa incluir outras ocorrências que determinem sintomas aéreos similares à infecção por nematoides, o resultado das classificações é consistente com os dados levantados em campo (como mostram as altas acurácias obtidas a partir dos dados de validação), assim como a quantidade de pixels atribuídos a cada classe nas diferentes datas de tomada da imagem. Nesse caso, pode-se verificar pela Tabela 18 que o mês de janeiro teve a menor área de café infectado, para ambos os classificadores e que o classificador RNA obteve as menores predominâncias da classe em relação ao FA para as três datas de aquisição das imagens. De fato, os sintomas da presença de nematoides tornam-se mais evidentes nos períodos mais secos (julho) e vão desde a murcha, amarelecimento, queda das folhas, redução do crescimento, podendo levar à morte da planta (MESQUITA et al., 2016).

De maneira geral, os resultados obtidos pelos indicadores estatísticos e pelo processo de classificação dos dados para as imagens referentes às três fases fenológicas do café foram consistentes e representaram adequadamente os dados levantados em campo. Esse resultado indica o potencial da região espectral do infravermelho de ondas curtas (SWIR) na detecção de nematoides em plantações de café, ou pelo menos de situações diversas no cafeeiro que apontam para uma necessidade de investigar melhor algumas áreas cultivadas, já que os sintomas podem confundir com outras ocorrências. Destaca-se também os índices de vegetação que foram utilizados com a substituição da banda do NIR pela do RE3 que mostraram grande potencial na discriminação das áreas saudáveis e infectadas.

7. CONCLUSÕES E RECOMENDAÇÕES

O desenvolvimento de técnicas de aquisição de dados a partir de sensores multiespectrais vem beneficiando várias ciências ao longo dos anos e trouxe, em especial com o sensoriamento remoto, importantes inovações no estudo e análise do uso e cobertura da terra. Essa tecnologia é capaz de agregar conhecimento aos estudos sobre cobertura vegetal, a partir da habilidade de discriminar alvos da superfície com base nas suas propriedades físico-químicas.

O acesso livre a dados de sensores com características inovadoras impulsiona a realização e a investigação de novos estudos. A configuração do sistema sensor Sentinel-2 combinada com abordagens de aprendizado de máquina, potencializa estudos em diversas aplicações, inclusive agrícolas.

Nesse contexto, o estudo sobre o potencial de diferentes índices de vegetação baseados na região do espectro do infravermelho de ondas curtas e demais atributos, combinados a abordagens baseadas em aprendizado de máquina na discriminação das áreas de café infectado e sadio foi desenvolvido na região do cerrado mineiro, mais precisamente na região cafeeira do município de Monte Carmelo – MG, assistida pela Emater-MG, utilizando dados do sensor MSI/Sentinel-2.

Para isso, inicialmente, foi feito um levantamento sobre os índices de vegetação mais promissores para o estudo encontrados na literatura, assim como dos aspectos teóricos da classificação baseada em aprendizado de máquina e dos algoritmos RNA e FA. Em seguida, foi apresentada a área de estudo e as características dos dados de referência utilizados, bem como das imagens multiespectrais MSI/Sentinel-2 e das principais ferramentas computacionais necessárias para o desenvolvimento do estudo. Por fim, foi descrita a metodologia utilizada para o desenvolvimento da pesquisa e os resultados obtidos.

No que se refere à seleção dos atributos mais adequados para a discriminação de áreas de café infectado por nematoides e áreas sadias, os resultados mostraram que, em relação às bandas do Sentinel-2, a região espectral do SWIR foi a que apresentou maior discriminação entre as classes, mostrando sua eficiência na discriminação espectral entre áreas de café infectado e sadio, com base nos dados levantados em campo. Além disso, o *Red Edge 3* também indicou um grande potencial. Em relação aos índices de vegetação, o experimento 1 indicou que o NDWI_RE3, NDVI_RE3, CSI_RE3 e MSI_RE3 resultaram em maior separabilidade entre as classes. Já o experimento 2, apontou para os

índices CSI_RE3, NBR2 e NDWI_RE3, nessa ordem, com maior potencial de contribuição na discriminação das classes. De modo geral, apesar do potencial claro da região espectral das bandas do SWIR e RE3, os índices de vegetação se mostraram mais favoráveis na detecção de áreas infectadas por nematoides no cultivo cafeeiro, para os dados levantados em campo para os dois experimentos realizados.

Com base nesses resultados, foram definidos 5 conjuntos de atributos que se julgou oportunos na detecção do problema e foi realizada uma comparação da aplicação desses conjuntos no treinamento dos classificadores FA e RNA, com base nas amostras coletadas dos dados do segundo levantamento de campo disponibilizados pela Universidade Federal de Uberlândia (UFU). Os resultados mostraram que o conjunto de melhor desempenho foi aquele composto por 11 atributos (conjunto 5): bandas do Vermelho, RE3, NIR, SWIR 1 e SWIR 2, imagens de textura baseadas nas bandas do NIR e SWIR 1, e os índices MSI_RE3, CSI_RE3, MCARI e TCARI. A aplicação dos algoritmos de classificação mostrou melhor desempenho para FA em relação à RNA, tanto nos indicadores estatísticos gerados, como no tempo de processamento que foi significativamente menor.

De maneira geral, os resultados mostrados pelos indicadores estatísticos com a aplicação dos dois classificadores utilizados, com base no conjunto de atributos de melhor desempenho, foram satisfatórios para as imagens referentes às três fases fenológicas do café. Esse resultado confirma o potencial da utilização da região espectral do infravermelho de ondas curtas (SWIR) na detecção de nematoides em plantações de café. Destaca-se também os índices que foram utilizados com a substituição da banda do NIR pela do RE3, que ofereceram melhores desempenhos em relação àqueles com as formulações originais.

Finalizando, foi verificada a capacidade das ferramentas baseadas em aprendizado de máquina no mapeamento de talhões de café infectados por nematoides, com base no conjunto de atributos mais adequado. Para isso, foram classificadas as cenas de julho de 2018 e janeiro e abril de 2019, referente a três fases do ciclo fenológico do café distintas.

Os mapas temáticos gerados apresentaram uma distribuição espacial das classes similar em ambos os algoritmos, indicando a robustez de métodos baseados em aprendizado de máquina. Apesar do resultado semelhante, as imagens de incerteza geradas a partir das classificações pela RNA, nas três datas de aquisição das cenas, definiram menores incertezas comparadas as de FA, embora esse último classificador

tenha definido melhores indicadores de desempenho. No que se refere à definição das áreas onde a atribuição de um pixel a uma classe foi mais incerta, a RNA teve desempenho superior ao FA. A predominância de talhões infectados nas três datas classificadas é claramente perceptível, mas há de se ressaltar que essa detecção pode incluir outras ocorrências que definem sintomas similares à infecção por nematoides.

O tema abordado nesse trabalho é de grande relevância, uma vez que, a detecção de plantas infectadas por nematoides é dispendiosa e demanda a coleta e análise laboratorial de um grande número de amostras de solos infectados sendo, muitas vezes, uma abordagem inviável para o produtor. O mapeamento pode trazer indícios de localizações com a ocorrência de nematoides no cafeeiro, se tornando uma ferramenta adequada para o fornecimento informações que possam aumentar a eficiência do manejo e alcançar as expectativas de produção agrícola.

Por fim, algumas recomendações para trabalhos futuros podem ser feitas, tais como, a investigação de todas as fases fenológicas do café, se possível, com imagens de um mesmo ano. Um levantamento mais amplo de índices de vegetação que considerem a região espectral do SWIR, para a investigação de áreas com a ocorrência de nematoides no cafeeiro e a inclusão de classes que contemplem outros tipos de ocorrência com sintomas similares à infecção por nematoides. Além disso, outras abordagens baseadas em aprendizado de máquina também podem ser testadas.

Para o processo de treinamento dos classificadores, uma recomendação seria a apresentação de amostras de treinamento e de validação em momentos distintos, evitando assim, problemas como autocorrelação espacial entre os pixels utilizados nos dois processos, o que não ocorre nos procedimentos especificados pelo manual do *software* WEKA, que é inserido um único arquivo de amostras e o próprio *software* seleciona aleatoriamente a proporção de amostras de treinamento e validação especificados pelo usuário.

Também seria interessante a incorporação de imagens de radar de abertura sintética (SAR - *Synthetic Aperture Radar*) pois permitem a possibilidade de extrair métricas de textura mais adequadas devido à característica do sinal de micro-ondas.

REFERÊNCIAS

BARNES, E.M.; CLARKE, T.R.; RICHARDS, S.E.; COLAIZZI, P.D.; HABERLAND, J.; KOSTRZEWSKI, M.; WALLER, P.; CHOI, C.; RILEY, E.; THOMPSON, T.; LASCANO, R.J.; LI, H.; MORAN, M.S. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. **In: Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN**, pp.16–19, 2000.

BEHMANN, J.; MAHLEIN, A. K.; RUMPF, T.; RÖMER, C.; PLÜMER, L. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. **Precision Agriculture**, v. 16, p. 239-260, 2015.

BELGIU, M., DRĂGUȚ, L. Random forest in remote sensing: a review of applications and future directions. **ISPRS J. Photogramm. Remote Sens.** 114, 24–31, 2016.

BENEDIKTSSON, J. A.; SWAIN, P. H.; ERSOY, O. K. **Neural network approaches versus statistical methods in classification of multisource remote sensing data.** IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Vol. 28, No. 4, 1990.

BREIMAN, L. **Random Forests.** Statistics Department - University of California, Berkeley, CA 94720, 2001.

BUSETTO, L.; CASTELEYN, S.; GRANELL, C.; PEPE, M.; BARBIERI, M.; CAMPOS-TABERNER, M.; CASA, R.; COLLIVIGNARELLI, F.; CONFALONIERI, R.; CREMA, A.; GARCÍA-HARO, F. J.; GATTI, L.; GITAS, I. Z.; GONZÁLEZ-PÉREZ, A.; GRAU-MUEDRA, G.; GUARNERI, T.; HOLECZ, F.; KATSANTONIS, D.; MINAKOU, C.; MIRALLES, I.; MOVEDI, E.; NUTINI, F.; PAGANI, V.; PALOMBO, A.; PAOLA, F. D.; PASCUCCI, S.; PIGNATTI, S.; RAMPINI, A.; RANGHETTI, L.; RICCIARDELLI, E.; ROMANO, F.; STAVRAKLOUDIS, D. G.; STROPPIANA, D.; VIGGIANO, M.; BOSCHETTI, M. Downstream Services for Rice Crop Monitoring in Europe: From Regional to Local Scale. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 10, No. 12, 2017.

CAMARGO, A. P.; CAMARGO, M. B. P. **Definição e esquematização das fases fenológicas do cafeeiro arábica nas condições tropicais do Brasil.** Agrometeorologia. Bragantia, Campinas, 60(1), 65-68, 2001.

CHEN, C.; LIAW, A.; BREIMAN, L. **Using Random Forest to Learn Imbalanced Data.** Statistics Department. University of California, Berkeley, 2004.

CHEN, T.Y.; TSE, T. H.; YU, Y. T. Proportional Sampling Strategy: A Compendium and some Insights. **Journal of Systems and Software.** 58(1):65-81, 2001.

Companhia Nacional de Abastecimento (CONAB). **Observatório Agrícola. Acompanhamento da safra brasileira – café.** V.5 - SAFRA 2018 - N.4 - Quarto levantamento. Dezembro de 2018.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing of Environment**, 37, 35–46. 1991.

CONGALTON, R.G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. CRC press, 2008.

DAUGHTRY, C.S.T., WALTHALL, C.L., KIM, M.S., DE COLSTOUN, E.B. and MCMURTREY, J.E. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. **Remote Sensing of Environment**, 74, pp. 229–239, 2000.

Empresa de Assistência Técnica e Extensão Rural do Estado de Minas Gerais (EMATER - MG). **Conhecendo a Cafeicultura de Minas Gerais**, 2018. Disponível em: <<http://www.emater.mg.gov.br/portal.do>>. Acesso em 20 jun. 2019.

EFRON, B.; TIBISHIRANI, R. J. **An introduction to the bootstrap**. New York: John Wiley & Sons, 1993.

Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG). **Aspectos técnicos dos nematoides parasitas do cafeeiro**, 2011. Disponível em: <http://www.sapc.embrapa.br/arquivos/consorcio/boletins_tecnicos/bt_aspectos_tecnicos_nematoides.pdf>. Acesso em: 10 fev. 2019.

ESA (European Space Agency). **Sentinel Online**. Disponível em: <<https://earth.esa.int/web/sentinel/>>. Acesso em: 28 de junho 2019.

ESTABROOKS, A.; JO, T.; JAPKOWICZ, N. A multiple resampling method for learning from imbalanced data sets. **Computational Intelligence**, 20, 18–36, 2004.

Estação em Tempo Real - **Sismet cooxupé**. Disponível em: <sismet.cooxupe.com.br>. Acesso em 10 jan. 2020.

EVANS, J.T., 1989. Photosynthesis and nitrogen relationships in the leaves of C3 plants. **Oecologia**, 78(1), 9-19, 1989.

FREEMAN, E. A.; MOISEN, G.G.; FRESCINO, T. S. Evaluating effectiveness of downsampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada. **Ecological Modelling**, 233, 1–10, 2012.

GAO, B. C. NDWI - A Normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v. 58, p.257- 266, 1996.

GERSTMANN, H.; MÖLLER, M; GLÄßER, C. Optimization of spectral indices and long-term separability analysis for classification of cereal crops using multi-spectral RapidEye imagery. **International Journal of Applied Earth Observation and Geoinformation** 52, 115–125, 2016.

GISLASON, P.O., BENEDIKTSSON, J.A., SVEINSSON, J.R. Random forests for land cover classification. **Pattern Recogn. Lett.** 27, 294–300, 2006.

GONZALEZ, R.C.; WOODS, R. E. **Processamento Digital de Imagens**. 3. ed. São Paulo: Pearson Prentice Hall, 2010.

GONZALO-MARTÍN, C.; LILLO-SAAVEDRA, M.; GARCÍA-PEDRERO, A.; LAGOS, O.; MENASALVAS, E. Daily Evapotranspiration Mapping Using Regression Random Forest Models. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 10, No. 12, 2017.

HABOUDANE, D., MILLER, J.R., TREMBLAY, N., ZARCO-TEJADA, P.J. and DEXTRAZE, L. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. **Remote Sensing of Environment**, 81, pp. 416–426, 2002.

HALL, M. **Correlation-based Feature Selection for Machine Learning**, Dept of Computer Science, University of Waikato, 1999. Disponível em: <<http://www.cs.waikato.ac.nz/~mhall/thesis.pdf>>. Acesso em: 28 de junho de 2019.

HALL, M.; FRANK, E.; HOLMES, G; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA Data Mining Software: An Update**. Department of Computer Science. University of Waikato. Hamilton, New Zealand, 2009.

HALL, M.A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. **IEEE T KNOWL DATA EN**, 15, 1437–1447, 2003.

HAN, J.; PEI, J.; KAMBER, M. **DataMining: Concepts and Techniques**. Amsterdam, The Netherlands, 2011.

HERRMANN, L.; KARNIELI, A.; BONFIL, D. J.; COHEN, Y. ALCHANATIS, V. SWIR-based spectral indices for assessing nitrogen content in potato fields. **International Journal of Remote Sensing** Vol. 31, No. 19, 5127–5143, 2010.

HUNT, E.R.; ROCK, B.N. Detection of changes in leaf-water content using near-infrared and middle-infrared reflectances. **Remote Sensing of Environment**, v. 30, n. 1, p. 43-54, 1989.

IBRAHIM, H. M. M.; AHMAD, E. M.; MARTÍNEZ-MEDINA, A., ALY, M. A.M. Effective approaches to study the plant-root knot nematode interaction. **Plant Physiology and Biochemistry** 141 - 332-342, 2019.

International Coffee Organization (ICO). **Dados históricos**. Disponível em: <http://www.ico.org/pt/new_historical_p.asp>. Acesso em: 12 fev. 2019.

Instituto Agrônômico, Centro de Café “Alcides Carvalho” – IAC. **A luta contra a doença causada pelos nematoides parasitos do café**. O Agrônômico, Campinas, 59(1), 2007. Disponível em: <http://www.iac.sp.gov.br/publicacoes/agronomico/pdf/v59_artigo12.pdf>. Acesso em 21 jun. 2019.

JENSEN, J. R. **Sensoriamento Remoto do Ambiente: Uma perspectiva em recursos terrestres**. Tradução: José Carlos N. Epiphanyo, et al. São José dos Campos, SP: Parêntese, 2009.

JENSEN, J. R. **Introductory Digital Image Processing – A Remote Sensing Perspective**. 2nd Edition, Peterson, *Interactive and Animated Cartography*, 1996.

JONES, J.T., HAEGEMAN, A., DANCHIN, E.G.J., GAUR, H.S., HELDER, J., JONES, M.G.K., KIKUCHI, T., MANZANILLA-LÓPEZ, R., PALOMARES-RIUS, J.E., WESEMAEL, W.M.L., Perry, R.N. Top 10 plant-parasitic nematodes in molecular plant pathology. **Mol. Plant Pathol.**14, 946–961, 2013.

KAUFMAN, Y.J.; REMER, L. Remote sensing of vegetation in the mid-IR: The 3.75 μm channels. **IEEE Geoscience and Remote Sensing Letters**. v 32, p. 672-683, 1994.

KEY, C.H.; BENSON, N.C. The Normalized Burn Ratio (NBR): A Landsat TM radiometric measure of burn severity. **U.S. Department of the Interior, Northern Rocky Mountain Science Centre**, 1999.

KIM, H.; YEOM, J., 2014. Effect of red-edge and texture features for object-based paddy rice crop classification using RapidEye multi-spectral satellite image data. **International Journal of Remote Sensing**, 35(19), 7046-7068.

LARY, D. J. Artificial Intelligence in Geoscience and Remote Sensing. **Joint Center for Earth Systems Technology (JCET) UMBC, NASA/GSFC**, United States, 2010.

LARY, D. J.; ALAVI, A. H.; GANDOMI, A. H.; WALKER, A. L. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3-10, 2016.

LAWRENCE, R.L., WOOD, S.D., SHELEY, R.L. Mapping invasive plants using hyperspectral imagery and Breiman cutler classifications (randomForest). **Remote Sens. Environ.** 100, 356–362, 2006.

LICZBINSKI, C. A. **Classificação de Dados Imagem em Alta Dimensionalidade, empregando Amostras Semi-Rotuladas e Estimadores para as Probabilidades a Priori**. 2007. 152p. Dissertação (Mestrado em Sensoriamento Remoto) – UFRGS. Porto Alegre, RS, 2007.

LINDGREN, D. T. **Land use planning and remote sensing**. Managing editor: V. Klemas, University of Delaware, DE. ISBN 90-247-3083-X, 1985.

LING, C.; LI, C. **Data mining for direct marketing problems and solutions**. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York, 1998.

LOUIS, J.; DEBAECKER, V.; PFLUG, B.; MAIN-KNORN, M.; BIENIARZ, J.; MUELLER-WILM, U.; CADAU, E.; GASCON, F. **SENTINEL-2 SEN2COR: L2A PROCESSOR FOR USERS**. Proc. ‘Living Planet Symposium 2016’, Prague, Czech Republic, (ESA SP-740), 2016.

MAIN-KNORN, M.; PFLUG, B.; LOUIS, J.; DEBAECKER, V.; MÜLLER-WILM, U.; GASCON, F. **Sen2Cor for Sentinel-2**. Event: SPIE Remote Sensing, Warsaw, Poland, 2017.

MALAVOLTA, E.; KLIEMANN, H. J. **Desordens nutricionais no Cerrado**. Piracicaba: Potafos, 1985.

MARTINS, G. D.; GALO, M. L. B. T.; VIEIRA, B. S. Detecting and Mapping Root-Knot Nematode Infection in Coffee Crop Using Remote Sensing Measurements. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 10, No. 12, 2017.

MAXWELL, A. E.; WARNER T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: an applied review. **International Journal of Remote Sensing**. ISSN: 0143-1161 (Print) 1366-5901, 2018.

MESQUITA, Carlos Magno de et al. Manual do café: distúrbios fisiológicos, pragas e doenças do cafeeiro (*Coffea arábica* L.). Belo Horizonte: EMATER-MG, 2016. 62 p. il.

MOREIRA, M. A. **Fundamentos do sensoriamento remoto e metodologias de aplicação**. 4ª Edição atualizada e ampliada – Viçosa, MG: Ed. UFV, 2012.

MUTANGA, O.; DUBE, T.; GALAL, O. Remote sensing of crop health for food security in Africa: Potentials and constraints. **Remote Sensing Applications: Society and Environment** 8, 231–239, 2017.

NOVO, E. M. L. M. **Sensoriamento Remoto: princípios e aplicações**. 4ª Edição. São Paulo: Blucher, 2010.

OLIVEIRA, C. M. G. ROSA, J. M. O. **Boletim Técnico - Nematoides Parasitos do Cafeeiro**. Instituto Biológico. n. 32. São Paulo, 2018. Disponível em: <http://www.biologico.agricultura.sp.gov.br/uploads/files/pdf/Boletins/cafe/nematoides_parasitos_cafeeiro.pdf>. 2019.

PONZONI, F. J.; SHIMABUKURO, Y. E.; KUPLICH, T. M. **Sensoriamento Remoto da Vegetação**. 2ª Edição. Atualizada e Ampliada. São Paulo: Oficina de Textos, 2012.

POWERS, D. M. W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. **Mach. Learn. Technol**, 2, 2008.

RICHARDS, J. A.; JIA, X. **Remote Sensing Digital Image Analysis: An Introduction**. 2006.

ROUSE, J.W.; HAAS, R.H.; SCHELL, J.A.; DEERING, D.W. Monitoring vegetation systems in the great plains with erts. **Third ERTS-1 Symposium NASA**, NASA SP-351, Washington DC, 309-317, 1974.

SCHUSTER, C.; FÖRSTER, M.; KLEINSCHMIT, B., 2012. Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data. **International Journal of Remote Sensing**, 33(17), 5583–5599.

SMITH, A.M.S.; DRAKE, N.A.; WOOSTER, M.J.; HUDAK A.T.; HOLDEN, Z.A.; GIBBONS, C.J. Production of Landsat ETM+ Reference Imagery of Burned Areas within Southern African Savannas: Comparison of Methods and Application to MODIS. **International Journal of Remote Sensing**, v. 28, n. 12. p. 2753-2775, 2007.

SMITH, A. M. S.; WOOSTER, M. J.; DRAKE, N. A.; DIPOTSO, F. M.; FALKOWSKI, M. J.; Hudak A. T. Testing the potential of multi-spectral remote sensing for retrospectively estimating fire severity in African savanna environments. **Remote Sensing of Environment**, 97, 92–115, 2005.

STUMPF, A.; KERLE, N. Object-Oriented Mapping of Landslides Using Random Forests. **Remote Sensing of Environment**. 115: 2564–2577, 2011.

SUN, H.; LIU, W.; WANG, Y.; YUAN, S. Evaluation of Typical Spectral Vegetation Indices for Drought Monitoring in Cropland of the North China Plain. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 10, No. 12, 2017.

VAPNIK, V. N. **Statistics for Engineering and Information Science – The Nature of Statistical Learning Theory**. Second Edition. ISBN: 0-387-98780-0, 1999.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern Recogn**, 44, 330–349, 2011.

ZHANG, H.; LI, Q.; LIU, J.; SHANG, J.; DU, X.; MCNAIRN, H.; CHAMPAGNE, C.; DONG, T.; LIU, M. Image Classification Using RapidEye Data: Integration of Spectral and Textual Features in a Random Forest Classifier. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Vol. 10, No. 12, 2017.

WEISS, G.M.; PROVOST, F. Learning when training data are costly: the effect of class distribution on tree induction. **Journal of Artificial Intelligence Research**, 19, 315–354, 2003.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**, 2005.