

AMÉLIA LAÍSY DO NASCIMENTO

**ESTIMATIVA DE PRODUTIVIDADE DE CAFÉ POR MEIO DE MÉTODOS DE
MACHINE LEARNING**

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2019

**Ficha catalográfica preparada pela Biblioteca Central da Universidade
Federal de Viçosa - Câmpus Viçosa**

T

N244e
2019 Nascimento, Amélia Laísy do, 1987-
Estimativa de produtividade de café por meio de métodos
de machine learning / Amélia Laísy do Nascimento. – Viçosa,
MG, 2019.

vii, 73 f. : il. (algumas color.) ; 29 cm.

Orientador: Daniel Marçal de Queiroz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Agricultura de precisão. 2. Café - Rendimento.
3. Imagens de sensoriamento remoto. 4. LANDSAT (Satélites).
5. Ciência de dados. 6. Sensoriamento remoto. 7. Inteligência
artificial. I. Universidade Federal de Viçosa. Departamento de
Engenharia Agrícola. Programa de Pós-Graduação em
Engenharia Agrícola. II. Título.

CDD 22. ed. 631.3

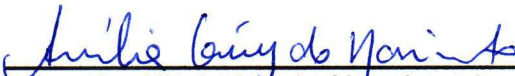
AMÉLIA LAÍSY DO NASCIMENTO

**ESTIMATIVA DE PRODUTIVIDADE DE CAFÉ POR MEIO DE MÉTODOS DE
MACHINE LEARNING**


Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 31 de julho de 2019.

Assentimento:



Amélia Laísy do Nascimento
Autora



Daniel Marçal de Queiroz
Orientador

AGRADECIMENTOS

A Deus por tudo que Ele tem proporcionado em minha vida.

A Universidade Federal de Viçosa e ao Departamento de Engenharia Agrícola (DEA) pela oportunidade de realizar o doutorado;

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e à Fundação de Amparo à Pesquisa do estado de Minas Gerais pelo apoio financeiro para realização desta pesquisa;

Aos professores Daniel Marçal de Queiroz, Domingos Sárvio Magalhães Valente, Francisco de Assis de Carvalho Pinto, Elpídio Inácio Fernandes Filho e Paulo Roberto Cecon pela atenção e pelos conselhos para desenvolvimento deste trabalho;

Ao Sr. Afonso por disponibilizar a Fazenda Braúna para realização do trabalho;

Aos professores Daniel Marçal de Queiroz, Domingos Sárvio Magalhães Valente, Elpídio Inácio Fernandes Filho, Flora Maria de Melo Villar e Williams Pinto Marques Ferreira pela participação na banca da defesa da tese;

Aos professores Lêda Rita D'Antonino Faroni e Daniel Marçal de Queiroz, coordenadores do Programa de Pós-graduação em Engenharia Agrícola da UFV enquanto eu fui discente do programa, e ao Délio Duarte e Rafael Assis Damasceno, secretários da Pós-graduação do DEA durante o período em que fui discente do programa em nível de mestrado e doutorado, pela atenção, cordialidade e pelo atendimento sempre que necessário;

Ao amigo e companheiro Emanuel por me ajudar, me aguentar e me proporcionar um lar em Viçosa;

A Arya que apareceu em minha vida em um momento de raiva, angústia e tristeza trazendo alegria, amor e felicidade ao meu coração e minha vida;

Aos amigos que conquistei por meio da Arya e que tornaram minhas noites em Viçosa muito divertidas;

Aos amigos e colegas do Laboratório de Mecanização Agrícola (LMA) pela convivência agradável, trocas de ideias e descontração;

A minha família e amigos que estão sempre comigo.

RESUMO

NASCIMENTO, Amélia Laisy do, D.Sc., Universidade Federal de Viçosa, julho de 2019. **Estimativa de produtividade de café por meio de métodos de machine learning**. Orientador: Daniel Marçal de Queiroz. Coorientador: Domingos Sárvio Magalhães Valente.

A produtividade agrícola representa o resultado de ações tomadas antes da colheita e indica se as práticas agrícolas adotadas causaram aumento ou redução no rendimento e podem ajudar na tomada de decisões futuras. Dessa maneira, a previsão de produtividade é uma ferramenta útil para os agricultores. Existem modelos que estimam a produtividade, porém, a quantidade de variáveis necessárias e a dificuldade em mensurá-las são um problema. Vários pesquisadores têm usado imagens orbitais para realizar estimativas de biomassa e produtividade de culturas. Além disso, alguns pesquisadores vêm combinando métodos de aprendizado de máquina (*machine learning*), mineração de dados (*data mining*) ou inteligência artificial (*artificial intelligence*) na tentativa de prever a produtividade de culturas agrícolas. Para estimar a produtividade agrícola, é interessante que o banco de dados possua imagens de todo o ciclo produtivo da cultura. Porém, o período de revisita dos satélites e a presença de nuvens sobre a área de estudo podem tornar o banco de dados incompleto. Uma possibilidade é adquirir imagens capturadas por sensores a bordo de distintos satélites. No entanto, cada sensor captura faixas de comprimento de onda diferentes e alguns sensores não capturam todos os comprimentos de onda necessários aos estudos. Uma forma de resolver esse problema é realizar uma predição das imagens faltantes de um satélite utilizando como base imagens oriundas de outro satélite. Dessa forma, consegue-se preencher lacunas na série de dados e garantir um banco de dados com uma série temporal mais representativa. Por fim, é possível utilizar a série temporal de informações derivadas das imagens orbitais para estimar a produtividade de culturas agrícolas. Portanto, o objetivo desta tese foi estimar a produtividade do café por meio de informações espectrais e *machine learning*. Para isso, o banco de dados foi composto por imagens Sentinel-2 originais, além de imagens Sentinel-2 preditas com base em imagens oriundas do Cbers-4, Landsat-8 e Resourcesat-2. A predição de imagens Sentinel-2 ocorreu por meio de sete métodos de *machine learning*. Os dados foram separados em conjunto de treinamento, teste e avaliação dos modelos. O desempenho dos modelos foi mensurado pela raiz do erro quadrático médio (*root*

mean square error - RMSE) entre o valor real e o valor predito pelo modelo para o conjunto que ficou de fora do treinamento. O teste t a 5% de significância foi usado para verificar a existência de igualdade ou diferença estatística entre os erros apresentados pelos modelos de predição da reflectância. Os métodos de *machine learning* mostraram-se eficazes para estimar os valores de reflectância de imagens Sentinel-2 com base em imagens oriundas do Cbers-4, do Landsat-8 e do Resourcesat-2. Os modelos que apresentaram menores RMSE's na predição da reflectância de imagens orbitais de uma data distinta a data cujos dados foram usados para treinar os modelos foram usados para estimar as imagens Sentinel-2 ausentes do banco de dados usado para estimar a produtividade do café. A partir das imagens orbitais, seis índices de vegetação e reflectância em seis bandas espectrais foram obtidos. A estimativa da produtividade ocorreu por meio de seis métodos de *machine learning*. Os modelos de estimativa foram implantados em linguagem R no programa computacional R Versão 3.5.1 (R Team, 2018). A raiz do erro quadrático médio (root mean square error - RMSE) e o erro médio absoluto (*mean absolute error* - MAE) foram usados para avaliar a acurácia dos modelos de estimativa da produtividade. O RMSE e o MAE serviram de entrada para o teste de Scott-Knott que agrupou os modelos semelhantes. Os métodos de *machine learning* apresentaram erros RMSE e MAE da estimativa da produtividade semelhantes uns aos outros pelo teste de Scott-Knott, com exceção da regressão linear utilizando 14 variáveis preditoras. Foi possível estimar a produtividade por meio de cinco variáveis com erros semelhantes aos erros apresentados pelos modelos com 10 e com 14 variáveis referentes a informações espectrais, topográficas e agronômicas. O erro RMSE mínimo apresentado pelos modelos correspondeu a uma diferença de 11% entre o valor estimado e o valor real da produtividade do café do talhão Pasto Novo 1 no ano de 2017. O erro MAE mínimo correspondeu a uma diferença de 1,7% entre o valor estimado e o valor observado da produtividade do talhão Açude 3 no ano de 2018. A estimativa da produtividade pode ser realizada com até três meses de antecedência.

ABSTRACT

NASCIMENTO, Amélia Laisy do, D.Sc., Universidade Federal de Viçosa, July, 2019. **Coffee productivity estimation through machine learning methods.** Advisor: Daniel Marçal de Queiroz. Co-advisor: Domingos Sárvio Magalhães Valente.

Yield represents the result of actions taken before harvest and indicates whether the agricultural practices adopted have caused yield increase or reduce, and may help in future decision making. In this way, the yield estimation is a useful tool for farmers. There are some models that try to estimate the yield, but the number of variables needed and the difficulty in measuring them is a problem. Several researchers have used orbital images to perform estimates of biomass and crop yields. In addition, some researchers have been combining machine learning, data mining or artificial intelligence methods in an attempt to estimate crop yields. To estimate the yield, it is interesting that the database has images of the entire crop's productive cycle. However, the satellite re-visit period and the presence of clouds over the study area may make the database incomplete. One possibility is to acquire images captured by sensors on board different satellites. However, each sensor captures different wavelength ranges and some sensors do not capture all the wavelengths required by the studies. One way to solve this problem is to predict the missing images of a satellite based on images from another satellite. In this way, it is possible to fill gaps in the data series and guarantee a database with a more representative time series. Finally, it is possible to use the time series of information derived from orbital images to estimate the yield of agricultural crops. Therefore, the purpose of this thesis was to estimate coffee yield through spectral information and machine learning. For this, the database was composed of original Sentinel-2 images, as well as Sentinel-2 images predicted based on images from Cbers-4, Landsat-8 and Resourcesat-2. The prediction of Sentinel-2 images occurred through seven methods of machine learning. The data were separated together in training, testing and evaluation of the models. Model performance was measured by the root mean square error (RMSE) between the actual value and the value predicted by the model for the set that was left out of training. The 5% significance t-test was used to verify the existence of statistical equality or difference between the errors presented by the reflectance prediction models. Machine learning methods proved to be effective in predicting the reflectance values of Sentinel-2 images based on Cbers-4, Landsat-8 and

Resourcesat-2 images. The models that presented lower RMSE's in predicting the reflectance of orbital images from a date other than the date whose data were used to train the models were used to predict Sentinel-2 images missing from the database used to estimate the coffee yield. From the orbital images, six vegetation indices and reflectance in six spectral bands were obtained. The estimation of yield occurred through six methods of machine learning. The estimation models were implemented in R language in the R Version 3.5.1 (R Team, 2018). Mean root error square (RMSE) and mean absolute error (MAE) were used to evaluate the accuracy of yield estimation models. RMSE and MAE served as input to the Scott-Knott test that grouped similar models. The machine learning methods presented RMSE and MAE errors of yield estimation similar to each other by the Scott-Knott test, with the exception of linear regression using 14 predictor variables. It was possible to estimate the yield through five variables with errors similar to the errors presented by the models with 10 and with 14 variables referring to spectral, topographic and agronomic information. The minimum RMSE error presented by the models corresponded to a difference of 11% between the estimated value and the real productivity value of the Pasto Novo 1 field coffee in 2017. The minimum MAE error corresponded to a difference of 1.7% between the estimated value and the observed yield value of the Açude 3 field in the year 2018. Productivity estimates can be made up to three months in advance.

SUMÁRIO

1	INTRODUÇÃO GERAL.....	1
1.1	Disposição do trabalho.....	3
1.2	Referências.....	3
2.	Predição de imagens orbitais por métodos de <i>machine learning</i> para complementação de série histórica de dados.....	5
2.1	INTRODUÇÃO.....	9
2.2	MATERIAL E MÉTODOS.....	11
2.3	RESULTADOS E DISCUSSÃO.....	17
2.4	CONCLUSÃO.....	37
2.5	REFERÊNCIAS.....	38
3.	Estimativa da produtividade por meio de perfis espectrais do cafeeiro e métodos de <i>machine learning</i>	41
3.1	INTRODUÇÃO.....	43
3.2	MATERIAL E MÉTODOS.....	45
3.2.1	Área de estudo.....	45
3.2.2	Imagens orbitais.....	45
3.2.3	Pré-processamento das imagens orbitais.....	46
3.2.4	Índices espectrais.....	49
3.2.5	Modelo digital de elevação.....	51
3.2.6	Produtividade da cultura.....	52
3.2.7	Análise dos dados e modelo para estimativa de produtividade.....	53
3.2.8	Correlação entre as variáveis preditoras.....	54
3.2.9	Treinamento dos modelos de estimativa da produtividade.....	54
3.2.10	Avaliação dos modelos de estimativa da produtividade.....	56
3.3	RESULTADOS E DISCUSSÃO.....	57
3.3.1	Produtividade de café.....	57
3.3.2	Correlação entre as variáveis preditoras.....	57
3.3.3	Importância das variáveis de entrada nos modelos de estimativa da produtividade de café.....	59
3.3.4	Teste dos modelos de estimativa da produtividade de café.....	61
3.3.5	Avaliação dos modelos de estimativa da produtividade de café.....	67
3.4	CONCLUSÃO.....	69
3.5	REFERÊNCIAS.....	69
4.	Conclusões Gerais.....	73

1 INTRODUÇÃO GERAL

A alta demanda do mercado por café é suprida por produtores de mais de 70 países, sendo que destes, apenas três fornecem mais de 50% do total produzido no mundo (FAO, 2015). Embora exista diferença na escala de produção entre os países e, também, entre os produtores, a produtividade da lavoura é uma preocupação para todos. De modo que a verificação de recursos materiais e financeiros, assim como o dimensionamento da infraestrutura e maquinários, além da contratação de mão-de-obra são atividades que dependem diretamente da produtividade da lavoura. Outro aspecto a ser destacado, consiste no fato de que a produtividade representa o resultado das ações realizadas antes da colheita e indica se as práticas agrícolas adotadas foram adequadas, o que pode auxiliar em tomadas de decisões futuras. Sendo assim, diante da importância que o conhecimento da produtividade tem no estabelecimento das técnicas de manejo a serem adotadas na lavoura, a estimativa da produtividade é uma ferramenta útil aos produtores.

A produtividade agrícola depende de muitas variáveis que geralmente são utilizadas como base em modelos matemáticos para sua estimativa, como os modelos agrometeorológicos (GOMES et al., 2014; NUNES et al., 2010). A resposta espectral da planta é resultado do seu vigor vegetativo, que por sua vez, relaciona-se à produtividade agrícola (PICOLI et al., 2009). Dessa forma, índices espectrais, como o Índice de Vegetação da Diferença Normalizada (*Normalized Difference Vegetation Index* – NDVI), passaram a ser incluídos na modelagem para predição da produtividade, dando origem aos modelos agrometeorológicos-espectrais (KRÜGER; FONTANA; MELO, 2007; ROSA et al., 2010).

O ideal é que os dados de entrada dos modelos de estimativa da produtividade sejam obtidos de maneira fácil. O problema dos modelos existentes para estimativa da produtividade é a quantidade de variáveis ou a dificuldade em mensurá-las. Os modelos agrometeorológicos e os agrometeorológicos-espectrais necessitam de informações de temperatura média do ar, velocidade do vento a 2 m de altura, umidade relativa do ar, insolação ou radiação solar, precipitação pluvial, altitude, capacidade de armazenamento de água no solo e profundidade do sistema radicular (ROSA et al., 2010).

Imagens orbitais têm sido utilizadas para estimativas de biomassa e produtividade de culturas (CAI et al, 2019; SHIU; CHUANG, 2019; PETERSEN, 2018; SONG et al., 2016; YUE; YANG; FENG, 2016; NIGAM et al., 2016). Diversos pesquisadores empregam métodos de aprendizado de máquina (*machine learning*), mineração de dados (*data mining*) ou inteligência artificial (*artificial intelligence*) na tentativa de prever a produtividade de culturas agrícolas (PATEL; PATEL, 2016; GANDHI; ARMSTRONG, 2016; RAMESH; VARDHAN, 2015; MEDAR; RAJPUROHIT, 2014). A união de informações derivadas de imagens orbitais e métodos de *machine learning* pode facilitar a predição da produtividade agrícola. Uma vez que, para isso, o usuário precisa adquirir as imagens, processá-las e ajustar os modelos de estimativa da produtividade.

Para estimar a produtividade agrícola, é interessante que o banco de dados possua imagens de todo o ciclo produtivo da cultura. Porém, o período de revisita dos satélites e a presença de nuvens sobre a área de estudo podem tornar o banco de dados incompleto. Uma possibilidade é adquirir imagens capturadas por sensores a bordo de distintos satélites. Nesse caso, o usuário das imagens deve se preocupar com as características individuais de cada satélite, principalmente o fato de que cada sensor captura faixas de comprimento de onda diferentes.

Outra peculiaridade a ser observada é a ausência da captura de determinadas faixas espectrais por alguns sensores. O sensor Câmera Multiespectral Regular (MUX) a bordo do satélite *China-Brazil Earth Resources Satellite* (Cbbers-4) não detecta as bandas espectrais correspondentes ao infravermelho de ondas curtas 1 e 2 (SW1 e SW2). Já o sensor *Linear Imaging Self-Scanner* (LISS-III) do Resourcesat-2 não captura imagem na banda espectral do azul e do infravermelho de ondas curtas 2 (SW2). A ausência dessas bandas em determinadas datas não permite o cálculo de índices como o *Normalized Difference Water Index* (NDWI), *Green Vegetation Index* (GVI) e o *Tasseled Cap Wetness Index* (WETNESS). Ou seja, a ausência de bandas espectrais pode comprometer a realização de um trabalho.

Uma forma de resolver o problema de captura de faixas espectrais diferentes ou ausentes é realizar uma predição das imagens faltantes de um satélite utilizando como base imagens oriundas de outro satélite. A seguir, basta obter informações derivadas das imagens como a reflectância nas diversas bandas espectrais ou índices espectrais. Dessa forma, consegue-se preencher lacunas na série de dados

e garantir um banco de dados com uma série temporal mais representativa. Por fim, consegue-se utilizar a série temporal de informações derivadas das imagens orbitais para estimar a produtividade de culturas agrícolas.

O objetivo geral com este trabalho foi o de desenvolver um sistema para estimativa de produtividade do café utilizando métodos de *machine learning* e índices espectrais. Para isso, os seguintes objetivos específicos foram buscados:

- Analisar se diferentes métodos de *machine learning*, bem como distintas formas de representar as variáveis de entrada e saída dos modelos, predizem imagens orbitais oriundas de um satélite com base em imagens capturadas por sensores de outros satélites.
- Testar se a combinação entre os perfis espectrais da cultura do cafeeiro, as características topográficas e agronômicas da cultura e distintos métodos de *machine learning* contribui para a melhoria da estimativa da produtividade de uma lavoura de café.

1.1 Disposição do trabalho

Esta tese foi organizada em dois capítulos, além de introdução e conclusão geral. O primeiro capítulo consiste na predição de imagens orbitais com base em imagens obtidas por diferentes satélites para a complementação do banco de dados que foi utilizado no segundo capítulo. O segundo capítulo consiste na estimativa da produtividade da cultura do cafeeiro por meio de imagens orbitais obtidas durante todo o ciclo produtivo da cultura.

1.2 Referências

CAI, Y.; GUAN, K.; LOBELL, D.; POTGIETER, A. B.; WANG, S.; PENG, J.; XU, T.; ASSENG, S.; ZHANG, Y.; YOU, L.; PENG, B. Integrating satellite and climate data to predict wheat yield in Australia using machine-learning approaches. **Agricultural and Forest Meteorology**, v. 274. 2019.

FAO. **FAO Statistical Pocketbook: Coffee 2015**. Roma, 2015.

GANDHI, N.; ARMSTRONG, L. Applying machine learning techniques to predict yield of rice in humid subtropical climatic zone of India. In: **Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on**. IEEE, 2016.

- GOMES, A. C. S.; ROBAINA, A. D.; PEITER, M. X.; SOARES, F. C.; PARIZI, A. R. C. Modelo para estimativa da produtividade para a cultura da soja. **Ciência Rural**, v. 44, n. 1, 2014.
- KRÜGER, C. A. M. B.; FONTANA, D. C.; MELO, R. W. Estimativa do rendimento de grãos da soja no Rio Grande do Sul usando um modelo agrometeorológicoespectral regionalizado. **Revista Brasileira de Agrometeorologia**, v. 15, n. 3, 2007.
- MEDAR, R. A.; RAJPUROHIT, V. S. A survey on machine learning techniques for crop yield prediction. **International Journal of Advance Research in Computer Science and Management Studies**, vol. 2, n. 9, 2014.
- NIGAM, R.; VYAS, S. S.; BHATTACHARYA, B. K.; OZA, M. P.; MANJUNATH, K. R. Retrieval of regional LAI over agricultural land from an Indian geostationary satellite and its application for crop yield estimation. **Journal of Spatial Science**, 2016.
- NUNES, F. L.; CAMARGO, M.; FAZUOLI, L. C.; ROLIM, G. S.; PEZZOPANE, J. R. M. Modelos agrometeorológicos de estimativa da duração do estágio floração-maturação para três cultivares de café arábica. **Bragantia**, v. 69, n. 4, 2010.
- PATEL, H.; PATEL, D. A Comparative Study on Various Machine learning Algorithms with Special Reference to Crop Yield Prediction. **Indian Journal of Science and Technology**, v. 9, n. 22, 2016.
- PETERSEN, L. Real-Time Prediction of Crop Yields From MODIS Relative Vegetation Health: A Continent-Wide Analysis of Africa. **Remote Sensing**, v. 10, n. 11, 2018.
- PICOLI, M. C. A.; RUDORFF, B. F. T.; RIZZI, R.; GIAROLLA, A. Índice de vegetação do sensor MODIS na estimativa da produtividade agrícola da cana-de-açúcar. **Bragantia**, v. 68, n. 3, 2009.
- RAMESH, D.; VARDHAN, B.V. Analysis of crop yield prediction using machine learning techniques. **International Journal of Research in Engineering and Technology**, vol. 4, n. 1, 2015.
- ROSA, V. G. C.; MOREIRA, M. A.; RUDORFF, B. F. T.; ADAMI, M. Estimativa da produtividade de café com base em um modelo agrometeorológico-espectral. **Pesquisa agropecuária brasileira**, v. 45, n. 12, 2010.
- SHIU, Y. S.; CHUANG, Y. C. Yield estimation of paddy rice based on satellite imagery: comparison of global and local regression models. **Remote Sensing**, v. 11, n. 2, 2019.
- SONG, R.; CHENG, T.; YAO, X.; TIAN, Y.; ZHU, Y.; CAO, W. Evaluation of Landsat-8 time series image stacks for predicting yield and yield components of winter wheat. In: **Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International**. IEEE, 2016.
- YUE, J.; YANG, G.; FENG, H. Comparative of remote sensing estimation models of winter wheat biomass based on random forest algorithm. **Transactions of the Chinese Society of Agricultural Engineering**, v. 32, n. 18, 2016.

2. Predição de imagens orbitais por métodos de *machine learning* para complementação de série histórica de dados

RESUMO

Imagens orbitais têm sido utilizadas para diversos fins agrícolas, civis e militares. Dependendo do período de revisita do satélite e/ou da ocorrência de nuvens sobre a área estudada informações relevantes podem ser perdidas e a série histórica de dados fica incompleta. As lacunas na série de dados podem ser preenchidas com a utilização de imagens oriundas de satélites distintos. Porém, cada sensor captura a luz em faixas de comprimentos de ondas diferentes e alguns sensores não capturam todas as faixas espectrais capturadas por outro sensor. Dessa forma, objetivou-se com este trabalho estimar a reflectância de imagens Sentinel-2 ausentes com base nas reflectâncias das bandas de imagens oriundas dos satélites Cbers-4, Landsat-8 e Resourcesat-2. As reflectâncias das imagens foram adquiridas em seis datas em que houve coincidência de passagem de pelo menos dois satélites imageando a cena. Foram identificados 78 alvos na cena. A reflectância mínima, média, mediana e máxima dos pixels contidos nos 78 alvos foi determinada. Sete métodos de *machine learning* foram usados para estimar a reflectância de imagens orbitais. Os dados foram separados em conjunto de treinamento, teste e avaliação dos modelos. O desempenho dos modelos foi mensurado pela raiz do erro quadrático médio (*root mean square error* - RMSE) entre o valor real e o valor predito pelo modelo para o conjunto que ficou de fora do treinamento. A representação pelos valores mínimos, médios, medianos e máximos dos alvos e o método de *machine learning* foram considerados, respectivamente, os fatores A e B de um experimento fatorial. O teste F a 5% de significância na análise de variância (*analysis of variance* - ANOVA) foi utilizado para indicar se os fatores A e B atuaram de maneira independente nos modelos ajustados. O desdobramento da ANOVA foi realizado nos casos em que o teste F foi não significativo. O teste t a 5% de significância foi usado para verificar a existência de igualdade ou diferença estatística entre os erros apresentados pelos modelos de predição da reflectância. O teste t foi aplicado nas situações em que a ANOVA indicou atuação de maneira independente dos dois fatores, bem como quando o desdobramento da ANOVA deu

significativo para algum nível de um dos dois fatores. Os métodos de *machine learning* mostraram-se eficazes para estimar os valores de reflectância de imagens Sentinel-2 com base em imagens oriundas do Cbers-4, do Landsat-8 e do Resourcesat-2. Os modelos que apresentaram menores RMSE's na predição da reflectância de imagens orbitais de uma data distinta a data cujos dados foram usados para treinar os modelos foram: processo Gaussiano (*Gaussian Process - gaussprLinear*) e árvore de classificação e regressão com agregação Bootstrap (*treebag*) para as predições com base em imagens Cbers-4; regressão linear (*lm*), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel- svmLinear1*) e *treebag* nas predições com base em imagens Resourcesat-2; e, regressão ridge Bayesiana (*Bayesian Ridge Regression - bridge*) nas predições com base em imagens Landsat-8. Houve casos em que a banda espectral capturada pelo sensor a bordo de um satélite predisse a mesma banda espectral oriunda de outro satélite. Também houve situações em que a banda que compôs a variável preditora mais importante não correspondeu a mesma banda predita.

Palavras-chave: reflectância ,satélites, Cbers-4, Landsat-8, Resourcesat-2, redes neurais.

ABSTRACT

Orbital images have been used for many agricultural, civil and military purposes. Depending on the satellite's re-visit period and / or the occurrence of clouds over the studied area, relevant information may be lost and the historical data series can be compromised. The gaps in the data series can be filled with the use of images from different satellites. However, each sensor captures light in different wavelength ranges and some sensors do not capture all the spectral bands captured by another sensor. Thus, the objective of this work was to predict the reflectance of absent Sentinel-2 images based on reflectance in the image bands from the Cbers-4, Landsat-8 and Resourcesat-2 satellites. The reflectances of the images were acquired in six dates which there was coincidence of passage of at least two satellites imaging the scene. 78 targets were identified in the scene. The minimum, mean, median and maximum reflectance of the pixels contained in the 78 targets was determined. Seven methods of machine learning were used to predict orbital image reflectance. The data were separated in training, testing and evaluation data sets. The model performance was measured by the root mean square error (RMSE) between the actual value and the value predicted by the model for the set that was left out of training. The representation of the minimum, medium, median and maximum values of the targets and the machine learning method were considered, respectively, the factors A and B of a factorial experiment. The 5% significance test in analysis of variance (ANOVA) was used to indicate whether factors A and B acted independently on the fitted models. The ANOVA unfolding was performed in cases which the F test was not significant. The 5% significance test was used to verify the existence of equality or statistical difference among the errors presented by the reflectance prediction models. The t-test was applied in situations where ANOVA indicated performance independently of the two factors, as well as when the ANOVA unfolding showed significant for some level of one of the two factors. Machine learning methods proved to be effective in predicting the reflectance values of Sentinel-2 images based on Cbers-4, Landsat-8 and Resourcesat-2 images. The models that presented smaller RMSE's in the prediction of the reflectance of orbital images of a date different from the date whose data were used to train the models were: Gaussian Process (Gaussian Process - gaussprLinear) and tree of

classification and regression with aggregation Bootstrap (treebag) for the predictions based on Cbers-4 images; linear regression (lm), support vector machines with linear support with Linear Kernel - svmLinear1) and treebag in predictions based on Resourcesat-2 images; and Bayesian ridge regression (Bayesian Ridge Regression - bridge) in predictions based on Landsat-8 images. There were cases where the spectral band captured by the sensor on board one satellite predicted the same spectral band from another satellite. There were also situations which the band that composed the most important predictor variable did not correspond to the same predicted band.

Key words: reflectivity of distinct satellite bands, Cbers-4, Landsat-8, Resourcesat-2, artificial neural networks, support vector machines, linear regression.

2.1 INTRODUÇÃO

Imagens orbitais têm sido utilizadas para diversos fins agrícolas. Dentre eles, são exemplos, o mapeamento e gerenciamento de recursos naturais, estudos envolvendo transformação da paisagem, o mapeamento de uso e cobertura das terras e o diagnóstico de doenças em plantações (DU et al., 2019; SHIMRAH et al., 2019; FISHER et al., 2018; MOSE; WESTERN; TYRRELL, 2018; AHMAD; GOPARAJU; QAYUM, 2017; GURU; SESHAN; BERA, 2017; HALLETT et al., 2017; MCCARTHY et al., 2017; SANHOUSE-GARCÍA et al., 2016; SANHOUSE-GARCÍA et al., 2017). As estimativas de biomassa e de produtividade de culturas também são atividades beneficiadas pela utilização de imagens orbitais (CAI et al, 2019; SHIU; CHUANG, 2019; PETERSEN, 2018; SONG et al., 2016; YUE; YANG; FENG, 2016; NIGAM et al., 2016).

Diversas plataformas de sensoriamento remoto orbital, como o *Land Remote Sensing Satellite (Landsat-8)*, *Satellite pour l'Observation de la Terre (Spot)*, *National Oceanic and Atmospheric Administration (NOAA)*, *Radarsat*, *China-Brazil Earth Resources Satellite (Cbbers-4)*, *Ikonos*, *Terra*, *Sentinel-2* e o *Quick-Bird*, fornecem imagens da Terra. As imagens geradas por seus sensores possuem características distintas, como a faixa imageada (13 km até 5000 km), a resolução espacial (tamanho do pixel da imagem, podendo variar de 61 cm até 45 km), a resolução temporal (tempo de revisita de 1 até 176 dias) e a resolução espectral (número de bandas que cada sensor consegue captar, variando de 1 até 36 bandas por sensor) (CEPSRM, 2017). Características como essas definem o custo que as imagens podem ter ao usuário. No entanto, algumas plataformas disponibilizam gratuitamente as imagens obtidas por seus sensores, como é o caso das plataformas Cbbers-4, Landsat-8, Sentinel-2 e Resourcesat-2.

Estudos que utilizam séries temporais de imagens orbitais podem se beneficiar com a disponibilidade gratuita de imagens promovida por diversas plataformas. Porém, dependendo do estudo a ser realizado, do período de revisita do satélite ou da ocorrência de nuvens sobre a área estudada, informações relevantes podem ser perdidas tornando o banco de dados incompleto. Para preencher lacunas na série de dados e garantir uma série temporal mais representativa, existe a possibilidade de montar um banco de dados cobrindo diversas datas usando imagens oriundas de satélites distintos. Nesse caso, o

usuário das imagens deve se preocupar com as características individuais de cada satélite principalmente em relação à faixa de comprimento de onda sensível pelos sensores que são diferentes em cada satélite.

O comprimento de onda da banda espectral correspondente ao infravermelho próximo (NIR) capturado pelo sensor a bordo do satélite Cbers-4 varia de 0,77 a 0,89 μm . O Sentinel-2 possui o menor tempo de revisita que é de cinco dias. Isto é, em um determinado período de tempo, a presença de imagens Sentinel-2 é maior do que as imagens oriundas dos demais satélites. O sensor a bordo do Sentinel-2 captura a mesma banda espectral na faixa de comprimento de onda de 0,78 até 0,90 μm , enquanto que no Resourcesat-2 essa faixa varia de 0,77 até 0,86 μm . Já o Landsat-8 captura a banda espectral referente ao NIR na faixa de comprimento de onda de 0,85 a 0,88 μm . O Landsat-8 é o que apresenta a imagem NIR com a faixa espectral mais estreita dentre os demais satélites apresentados.

Além de diferentes faixas de comprimento de onda detectadas pelos sensores a bordo dos satélites, ocorre ainda a ausência de captura de algumas bandas espectrais por alguns sensores. O sensor Câmera Multiespectral Regular (MUX) a bordo do Cbers-4 e o sensor *Linear Imaging Self-Scanner* (LISS-III) do Resourcesat-2 não capturam, por exemplo, as bandas referentes ao infravermelho de ondas curtas 2 (SWIR 2). Caso exista a necessidade de uso de bandas espectrais que um dos satélites não captura, o uso de tal satélite pode comprometer a realização do trabalho.

Uma forma de utilizar imagens orbitais de mais de um satélite, contornando a problemática das faixas de comprimentos de onda diferentes, é prever as imagens ausentes de um satélite com base em imagens oriundas de outro satélite. Essa abordagem foi adotada por Filgueiras et al. (2019). Diante da ausência de bandas espectrais referentes ao infravermelho termal na captura de imagens pelos sensores a bordo dos satélites Sentinel-2, os autores utilizaram aprendizado de máquina e a temperatura da superfície do sensor TIRS (Sensor infravermelho termal) da plataforma Landsat-8 como referência para estimar a temperatura da superfície dos dados do sensor MSI/Sentinel-2. Os autores concluíram que a melhor estimativa da temperatura de superfície foi encontrada usando o algoritmo de floresta aleatória (*random forest*).

A predição de imagens ausentes de um satélite pode ser feita por meio de vários algoritmos de *machine learning*. Dessa forma, é possível compor um banco

de dados com imagens de satélites distintos, pois se consegue estimar os valores da imagem ausente a partir de imagens oriundas de outros satélites. Inclusive, tais predições podem ocorrer mesmo quando não houver alguma das bandas espectrais. Assim, objetivou-se com esse trabalho prever valores de reflectância nas bandas espectrais de imagens Sentinel-2 com base em dados de reflectância nas bandas de imagens oriundas dos satélites Cbers-4, Landsat-8 e Resourcesat-2.

2.2 MATERIAL E MÉTODOS

As imagens oriundas dos satélites Cbers-4 e Resourcesat-2 foram adquiridas junto à Divisão de Geração de Imagens do Instituto Nacional de Pesquisas Espaciais (INPE) enquanto as imagens do Landsat-8 foram obtidas junto ao *United States Geological Survey* (USGS). Já as imagens do Sentinel-2 foram adquiridas a partir do *Copernicus Open Access Hub*. As cenas de cada satélite englobaram a área contida na path/row correspondente a 217/074 do *Worldwide Reference System 2* (WRS-2).

Cada imagem adquirida passou por etapas de pré-processamento. A primeira delas foi referente ao georreferenciamento das imagens. Após essa etapa, a calibração radiométrica foi realizada utilizando a Equação 2.1. Assim, os números digitais das imagens foram transformados em valores de radiância. Utilizando a Equação 2.2, os valores de radiância foram convertidos em valores de reflectância. Na calibração das imagens oriundas do Landsat-8 e do Sentinel-2 foram utilizados coeficientes de calibração fornecidos nos metadados das imagens. Para a calibração das imagens do Resourcesat-2, o ganho do coeficiente da banda λ foi calculado por meio da Equação 2.3. Os dados de radiância máxima e mínima do sensor a bordo do Resourcesat-2 foram obtidos nos metadados da imagem. Já os dados de radiância máxima e mínima do sensor a bordo do Cbers-4, além do coeficiente de ganho da banda λ (G_λ) e do coeficiente de viés para a banda λ ($offset_\lambda$) foram obtidos nos trabalhos de Epiphanyo (2009) e Pinto et al (2016). A irradiância solar exoatmosférica nos sensores a bordo do Cbers-4 e do Resourcesat-2 foram obtidos nos trabalhos de Pinto et al (2016) e Keerthi e Kumar (2011). Após transformação dos níveis digitais das imagens em reflectância, foi realizada a correção atmosférica das imagens utilizando o método *Dark Object Subtraction* (DOS).

$$L_{\lambda} = G_{\lambda} \cdot DN_{\lambda} + offset_{\lambda} \quad \text{Eq. 2.1}$$

em que: L_{λ} é a radiância na banda λ no topo da atmosfera (*top of atmosphere* - TOA - $\text{W.m}^{-2}.\text{sr}^{-1}.\mu\text{m}^{-1}$); G_{λ} é o coeficiente de ganho da banda λ ($\text{W.m}^{-2}.\text{sr}^{-1}.\mu\text{m}^{-1}$); DN_{λ} é o número digital dos pixels da imagem e $offset_{\lambda}$ é o coeficiente de viés para a banda λ ($\text{W.m}^{-2}.\text{sr}^{-1}.\mu\text{m}^{-1}$).

$$\rho_{\lambda} = \frac{\pi L_{\lambda} d^2}{E_{SUN_{\lambda}} \cos \theta_z} \quad \text{Eq. 2.2}$$

em que: ρ_{λ} é a reflectância TOA na banda λ (adimensional); π é a constante matemática (adimensional); d é a distância Terra-Sol (unidades astronômicas); $E_{SUN_{\lambda}}$ é a irradiância solar exoatmosférica ($\text{W.m}^{-2}.\text{m}^{-1}$) e θ_z é o ângulo do zênite solar (radianos).

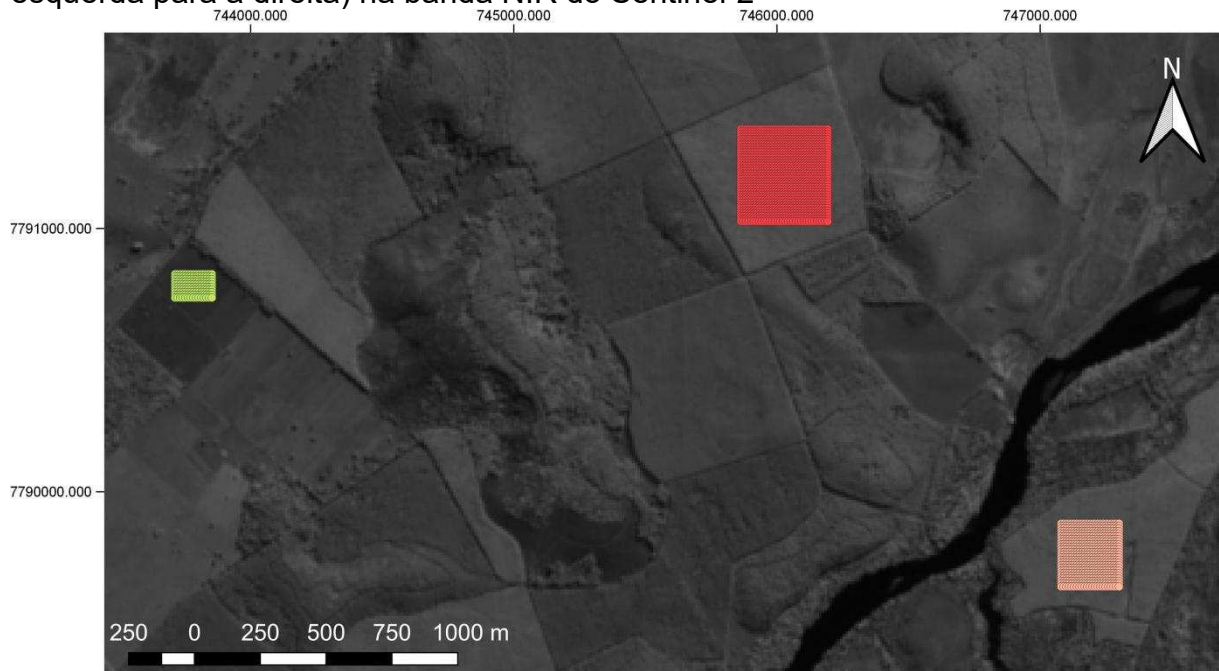
$$G_{\lambda} = \frac{L_{\max_{\lambda}} - L_{\min_{\lambda}}}{Q_{cal_{\max_{\lambda}}} - Q_{cal_{\min_{\lambda}}}} \quad \text{Eq. 2.3}$$

em que: $L_{\max_{\lambda}}$ é a radiância máxima da banda λ ($\text{W.m}^{-2}.\text{sr}^{-1}.\mu\text{m}^{-1}$); $L_{\min_{\lambda}}$ é a radiância mínima da banda λ ($\text{W.m}^{-2}.\text{sr}^{-1}.\mu\text{m}^{-1}$); $Q_{cal_{\max_{\lambda}}}$ é a quantização máxima na banda λ e $Q_{cal_{\min_{\lambda}}}$ é a quantização mínima na banda λ .

Após o cálculo da reflectância e correção atmosférica das imagens, foram escolhidos manualmente 78 alvos em comum às imagens capturadas pelos sensores a bordo dos satélites Cbers-4, Landsat-8, Resourcesat-2 e Sentinel-2. Os alvos foram edifícios, corpos hídricos, áreas com distintas formações vegetais e áreas com o solo nu. Após percepção visual, notou-se que alguns alvos possuíam tamanho adequado por conterem uma quantidade suficiente de pontos para o cálculo do valor médio. Porém em outros casos, o alvo foi bem delimitado pela imagem Sentinel-2 que possui 10 metros de resolução, mas não foi bem delimitado

pelas imagens com resoluções menores. Em casos como esse, nas imagens com menores resoluções poderia haver contaminação do pixel com a reflectância de áreas ao redor do alvo de interesse. Com isso, optou-se por representar os alvos por meio do valor mínimo, médio, mediano e máximo (Figura 2.1).

Figura 2.1. Recorte de cena do Sentinel-2 ilustrando os alvos números 54, 55, 56 (da esquerda para a direita) na banda NIR do Sentinel-2



O banco de dados foi composto pela reflectância de cada um dos 78 alvos nas bandas espectrais fornecidas pelos satélites Cbers-4, Resourcesat-2, Landsat-8 e Sentinel-2. Em cada data analisada houve alvos que ficaram encobertos por nuvens, logo, para cada data houve uma quantidade diferente de alvos sendo considerada nos modelos (Tabela 2.1). A banda Cirrus do Landsat-8 e do Sentinel-2 foi utilizada para ajudar a escolher alvos em que não havia nuvens os encobrindo. Foram selecionadas seis datas em que houve coincidência de passagem de pelo menos dois satélites imageando a cena (Tabela 2.1). Isto é, pelo menos dois satélites passaram imageando a cena no mesmo dia e horário. Em duas datas, expostas na Tabela 2.1, foram obtidas reflectâncias nos alvos para treinar os modelos, e em quatro datas as reflectâncias nos alvos foram usadas para avaliar os modelos de predição.

Tabela 2.1. Datas e quantidade de alvos a compor os bancos de dados de treinamento, teste e avaliação dos modelos

Satélite – Sensor	Treinamento e teste		Avaliação	
	Data	Quantidade de alvos	Data	Quantidade de alvos
Resourcesat-2 - LISS 3	12/08/2018	57	29/05/2018 e 03/06/2018	13
Landsat-8– OLI	23/06/2016	38	12/08/2018	40
Cbers-4 - MUX	23/06/2016	36 (35 na banda do azul)	11/09/2016	49

As etapas que envolveram manipulação das imagens como recortes, reprojeções e escolha dos alvos foram realizadas no *software* QGIS (QGIS Development Team, Open Source Geospatial Foundation, Chicago, IL, EUA). O pacote *Caret (Classification And Regression Training)* foi utilizado no processo de criação, teste e avaliação dos modelos. O *software* R Versão 3.5.1 (R Team, 2018) foi usado nas etapas referentes ao treinamento e avaliação dos modelos preditivos além das análises estatísticas e visualização dos resultados.

Os métodos de regressão linear (*lm*), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel - svmLinear1*), redes neurais artificiais (*nnet1*), árvore de classificação e regressão (*Classification and Regression Trees - CART*), CART com agregação Bootstrap (*Bootstrap Aggregating - bagged CART*), regressão ridge Bayesiana (*Bayesian Ridge Regression*) e processo Gaussiano (*Gaussian Process - gaussprLinear*) foram usados para estimar a reflectância nas bandas das imagens Sentinel-2. As bibliotecas e os parâmetros de ajuste utilizados para implementar os métodos são informadas na Tabela 2.2. O método *svmLinear1* utilizou o valor de um para o parâmetro custo e a rede neural artificial foi modelada a partir do uso de um neurônio na camada intermediária da rede. O parâmetro decaimento do peso usado na rede neural artificial foi igual a um.

Tabela 2.2. Técnicas, bibliotecas e parâmetros utilizados nos modelos

Nome	Representação do método	Biblioteca	Parâmetros de ajuste
Regressão linear	lm	nativa do R	
máquinas de vetores de suporte com Kernel linear (<i>support vector machine with linear Kernel</i>)	svmLinear1	e1071	custo = 1
rede neural artificial	nnet1	nnet1	número de neurônios na camada intermediária da rede = 1; decaimento do peso = 1
árvore de classificação e regressão (<i>Classification and Regression Trees - CART</i>)	rpart	rpart	
CART com agregação Bootstrap (<i>Bootstrap Aggregating - bagged CART</i>)	treebag	plyr ou e1071	
<i>Bayesian Ridge Regression</i>	bridge	monomvn	
<i>Gaussian Process</i>	gaussprLinear	caret	

Duas maneiras de controlar a aleatoriedade envolvida no ajuste dos modelos para estimar a reflectância nas bandas das imagens Sentinel-2 foram utilizadas para garantir resultados reproduzíveis. Primeiro, a função `set.seed` igual a um foi utilizada imediatamente antes da função de treinamento. A segunda maneira foi a criação de um objeto de controle que garantiu que todas as modelagens seguissem a mesma amostragem. O objeto de controle foi criado a partir da utilização do método de validação cruzada *k-fold*, o qual dividiu aleatoriamente os dados de treinamento em dez subconjuntos (10 folds). Nove desses subconjuntos foram reunidos e utilizados para treinamento enquanto o outro subconjunto foi usado para testar a predição. De maneira iterativa, cada subconjunto foi utilizado uma vez como conjunto de teste. Esse objeto de controle foi utilizado no treinamento de todos os modelos. Dessa

forma, em cada método de modelagem foram utilizados os mesmos dados de treinamento e teste. Com um subconjunto reservado para teste, o algoritmo calculou a raiz do erro quadrático médio (*root mean square error* - RMSE) entre o valor real e o valor predito pelo modelo para o subconjunto que não foi considerado no treinamento.

O método de *machine learning* e a forma de representação dos alvos foram considerados, respectivamente, os fatores A e B de um experimento fatorial. O fator A possuía sete níveis denominados de *lm*, *bridge*, *rpart*, *svmLinear1*, *nnet1*, *gaussprLinear* e *treebag*. O fator B possuía quatro níveis denominados *Min*, *Mean*, *Median* e *Max* que são referentes, respectivamente, à representação dos alvos por meio da reflectância mínima, média, mediana e máxima. A realização de uma reamostragem dos dados por meio da função *resamples* executada no programa R gerou o que foi considerado como as repetições do experimento.

Para testar se a interação entre os fatores foi significativa ou se os fatores atuaram de maneira independente, foi realizada uma análise de variância (*analysis of variance* – ANOVA). Isto é, foi verificado se os fatores A e B atuaram de maneira independente ou dependente sobre o RMSE resultante dos ajustes dos modelos de predição (hipótese nula). Os quadros da ANOVA foram construídos para as predições de cada banda predita com base em imagens oriundas de cada satélite. O teste F, a 5% de significância, foi utilizado para determinar a aceitação ou rejeição da hipótese nula. Nos casos em que a hipótese nula foi rejeitada, o desdobramento da ANOVA foi realizado.

Após a reamostragem pela função *resamples*, as diferenças entre os RMSE's resultantes de cada modelagem foram calculadas por meio do uso da função *diff* executada no programa R. O teste t, a 5% de significância, foi utilizado como parâmetro da função *diff* para verificar se as diferenças calculadas foram estatisticamente significativas. Essa análise foi realizada nas situações em que a ANOVA indicou atuação de maneira independente dos dois fatores, bem como quando o desdobramento da ANOVA foi significativo para algum nível de um dos dois fatores.

A combinação da forma de representar os alvos com o método de *machine learning* que ofereceu o RMSE mínimo resultante da reamostragem foi considerada como o melhor modelo para predição da reflectância nas bandas das imagens Sentinel-2. Essa conclusão foi estendida aos modelos cujos RMSE's foram

considerados estatisticamente iguais ao RMSE mínimo. Essa análise foi feita para as predições de cada banda do Sentinel-2 em função de cada satélite. Em seguida, dentre os modelos considerados como melhores, foram verificados quais modelos foram recorrentes nas predições das seis bandas do Sentinel-2.

Os modelos foram denominados pela união da forma de representar os alvos em determinada banda espectral da imagem (Min, Mean, Median ou Max) concatenado a um ponto(.) seguido pela nomenclatura dos métodos de *machine learning* (lm, svmLinear1, nnet1, rpart, treebag, bridge ou gaussprLinear). Dessa forma, o modelo denominado Max.lm, por exemplo, prediz a reflectância das bandas do Sentinel-2, cujos valores preditos corresponderam a máxima reflectância nos alvos da imagem, por meio do método de regressão linear (lm). De maneira similar, o modelo denominado Mean.nnet1 corresponde ao modelo que realizou a predição por meio do método de rede neural artificial (nnet1) e a banda predita foi representada pelos valores médios dos alvos.

O conjunto de dados para avaliação dos modelos foi utilizado com a finalidade de averiguar o RMSE da predição cujo treinamento utilizou reflectâncias de uma data para estimar a reflectância das imagens de outra data. O RMSE resultante desta etapa foi denominado de RMSE de referência. Essa análise também foi útil para verificar a importância de cada variável preditora. Para isso, os modelos passaram por um novo treinamento após retirada de uma variável preditora. Esse procedimento foi repetido e a cada iteração apenas uma variável era removida do conjunto de treinamento até que todas as variáveis predictoras ficaram ausentes no conjunto de treinamento uma vez. Cada iteração resultou em um novo RMSE. As diferenças entre esse RMSE e o RMSE de referência foram listadas em ordem decrescente. A variável cuja ausência no conjunto de treinamento gerou a maior diferença no RMSE foi considerada a mais importante.

2.3 RESULTADOS E DISCUSSÃO

A Tabela 2.3 e a Tabela 2.4 ilustram as ocasiões em que os fatores forma de representação dos alvos e o fator método de *machine learning* atuaram de maneira dependente (teste F significativo) sobre o RMSE resultante dos ajustes dos modelos de predição de cada banda com base em imagens oriundas de cada satélite. A

diferença entre as tabelas é que a Tabela 2.3 mostra os casos em que o Teste F foi significativo ou não quanto ao comportamento do fator A dentro do fator B. Já a Tabela 2.4 apresenta os casos em que o Teste F foi significativo ou não quanto ao comportamento de B dentro de A. Uma forma de interpretar o Teste F significativo é o entendimento de que ao menos um dos níveis de um fator foi afetado pela atuação do outro fator.

Tabela 2.3. Alteração do comportamento do fator forma de representação dos alvos (fator B) no fator métodos de *machine learning* (fator A) sobre o RMSE resultante dos ajustes dos modelos de predição de cada banda com base em imagens oriundas de cada satélite

Na predição das bandas:		com base em imagens:		bridge	gaussprLinear	lm	nnet1	rpart	svmLinear1	trebag
				Azul	com base em imagens:	Cbers-4	sim	sim	sim	não
Verde	sim	sim	sim	não			sim	sim	sim	
Vermelho	sim	sim	sim	não			sim	sim	sim	
NIR	não	não	não	não			não	não	não	
SW1	não	não	não	não			não	não	não	
SW2	sim	sim	sim	não			sim	sim	sim	
Azul	Landsat-8	não	não	sim		não	não	não	não	
Verde		não	não	sim		não	não	não	não	
Vermelho		não	não	sim		não	não	não	não	
NIR		não	não	sim		não	não	não	não	
SW1		não	não	sim		não	não	não	não	
SW2		não	não	sim		não	não	não	não	
Azul	Resourcesat-2	sim	sim	sim		não	sim	sim	sim	
Verde		sim	sim	sim		sim	sim	sim	sim	
Vermelho		sim	não	sim		sim	sim	não	sim	
NIR		não	não	não		não	não	não	não	
SW1		não	não	sim		não	não	não	não	
SW2		não	não	não		não	não	não	não	

Os métodos de *machine learning* são: regressão linear (lm), máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1), rede neural artificial com um neurônio na camada intermediária (nnet1), árvore de classificação e regressão (rpart), árvore de classificação e regressão com agregação Bootstrap (trebag), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente. A palavra sim indica que os métodos de mineração de dados (fator A) tiveram interação significativa com os níveis do fator

forma de representação dos alvos (fator B). A palavra não indica que os métodos de mineração de dados (fator A) não tiveram interação significativa com os níveis do fator forma de representação dos alvos (fator B).

Tabela 2.4. Alteração do comportamento do fator métodos de *machine learning* (fator A) no fator forma de representação dos alvos (fator B) sobre o RMSE resultante dos ajustes dos modelos de predição de cada banda com base em imagens oriundas de cada satélite

				Min	Mean	Median	Max
Na predição das bandas:	Azul	com base em imagens:	Cbbers-4	sim	sim	sim	sim
	Verde			sim	sim	sim	sim
	Vermelho			sim	sim	sim	sim
	NIR			não	não	não	não
	SW1			não	não	não	não
	SW2			sim	sim	sim	sim
	Azul		Landsat-8	sim	sim	sim	sim
	Verde			não	não	não	sim
	Vermelho			sim	sim	sim	sim
	NIR			sim	sim	sim	sim
	SW1			sim	sim	sim	sim
	SW2			sim	sim	sim	sim
Azul	Resourcesat-2	sim	sim	sim	sim		
Verde		sim	sim	sim	sim		
Vermelho		sim	sim	sim	sim		
NIR		não	não	não	não		
SW1		sim	sim	sim	sim		
SW2		não	não	não	não		

Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente. A palavra sim indica que as formas de representação dos alvos (fator B) tiveram interação significativa com os níveis do fator métodos de mineração de dados (fator A). A palavra não indica que as formas de representação dos alvos (fator B) não tiveram interação significativa com os níveis do fator métodos de mineração de dados (fator A).

A ausência de interação, teste F não significativo, apresentados na Tabela 2.3, apontou que a diferença no RMSE entre os níveis do fator A não dependeram dos

níveis do fator B. Já a ausência de interação, teste F não significativo, apresentados na Tabela 2.4, indicou que a diferença no RMSE entre os níveis do fator B não dependeram dos níveis do fator A. Nas predições das reflectâncias das bandas do infravermelho próximo (NIR), do infravermelho de ondas curtas um (SWIR 1) e do infravermelho de ondas curtas dois (SWIR 2) não houve interação entre os fatores A e B (Tabela 2.3 e Tabela 2.4). A predição do NIR em que isso ocorreu foi com base em imagens oriundas do Cbers-4 e também do Resourcesat-2. Já a predição do SWIR 1 e do SWIR 2 foi feita, respectivamente, com base em imagens Cbers-4 e Resourcesat-2. Esse resultado indicou que a forma de representação dos alvos não afetou os resultados obtidos na predição dessas bandas com a utilização de distintos métodos de *machine learning* (Tabela 2.3). O contrário também é verdadeiro (Tabela 2.4). Isto é, os métodos de *machine learning* não interferiram nos resultados obtidos na predição dessas bandas com a utilização de diferentes formas de representação dos alvos.

No quesito interação entre os fatores, o método de redes neurais artificiais (nnet1) se destacou dos demais. De 18 análises de interação, apenas em duas delas houve influência do fator forma de representação dos alvos (Teste F significativo - Tabela 2.3) quando a rede neural foi treinada com um neurônio na camada intermediária (nnet1). Os demais métodos apresentaram entre seis e sete análises cuja influência da forma de representar os alvos foi significativa. O destaque contrário ficou por conta do método lm que, dentre 18 análises de interação possíveis, 14 delas mostraram que houve influência da forma de representação dos alvos nos RMSE's obtidos nas modelagens. Esse resultado evidencia o poder de generalização das redes neurais já apontadas por outros autores (LEAL et al., 2015; OLIVEIRA et al., 2010; BINOTI; BINOTI; LEITE, 2014).

Em relação à atuação dos métodos de *machine learning* dentro dos níveis do fator forma de representação dos alvos foi observado que na maioria das análises de interação nos modelos de predição houve influência do método de mineração. Isto é, o método de *machine learning* fez diferença nos resultados obtidos na maioria dos modelos de predição. A exceção ocorreu nas predições das bandas NIR, SWIR 1 e SWIR 2 com base em imagens Cbers-4 e Resourcesat-2 - já mencionadas - e na predição da banda verde com base em imagens Landsat-8 (Tabela 2.4). Na predição da banda verde com base em imagens Landsat-8, os métodos de mineração não causaram diferença no RMSE quando os alvos foram representados pelo valor

mínimo, médio e mediano.

Os resultados do Teste F não foram conclusivos sobre a igualdade (ou diferença) estatística dos RMSE's médios resultantes dos modelos de predição estudados. Afinal, o Teste F significativo indicou a existência de ao menos um RMSE médio diferente dos demais. Como os fatores analisados possuíam mais de dois níveis, apenas o teste t foi capaz de indicar a existência (ou não) de igualdade entre os RMSE's apresentados pelos modelos.

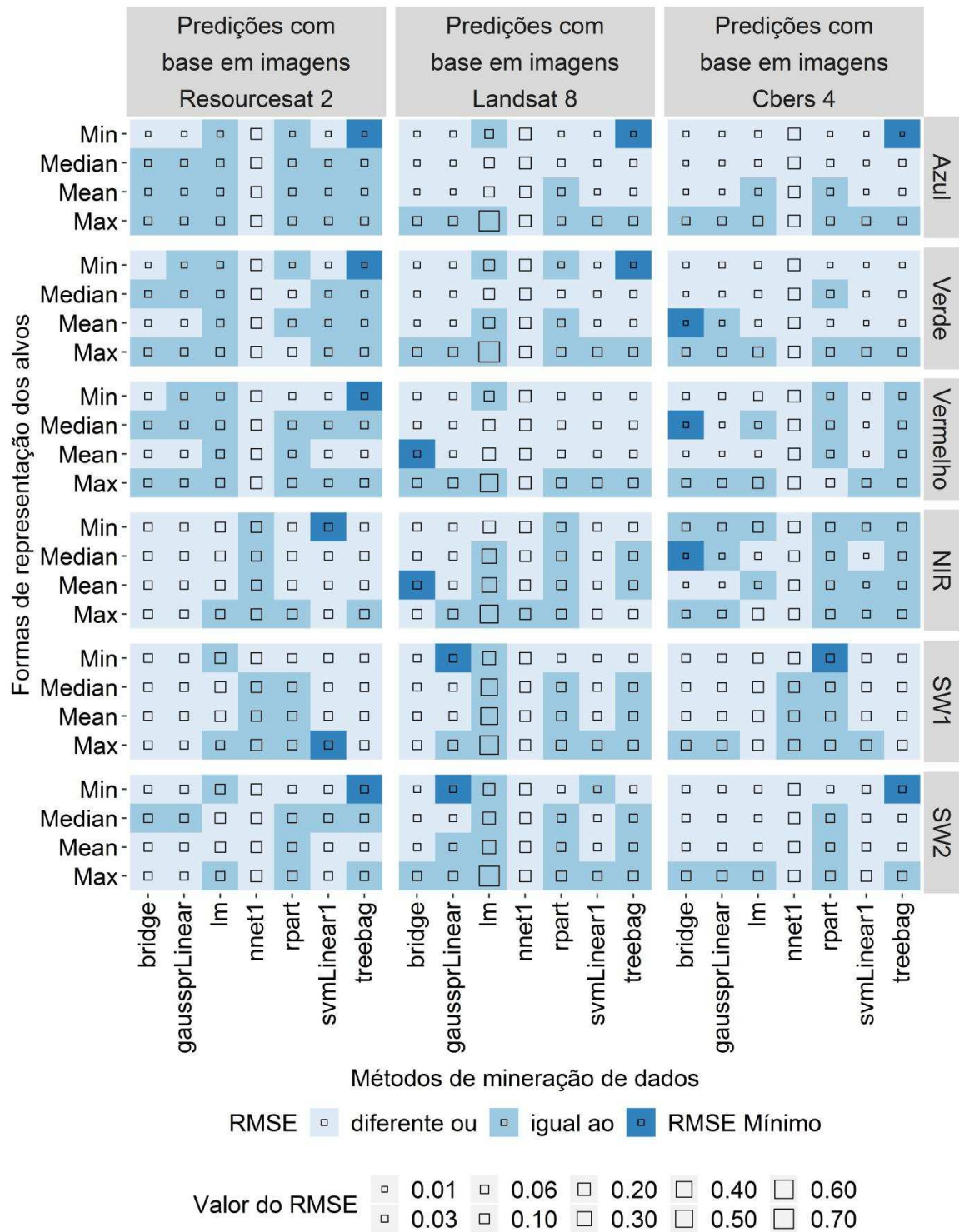
Embora a partir das análises com o Teste F não seja possível concluir qual método de *machine learning* e forma de representação dos alvos ofereceram os menores ou maiores RMSE's, seus resultados ainda podem ser úteis. No momento de caracterizar os alvos em uma cena, pode-se utilizar representações distintas (valores mínimos, máximos, médios, medianos, ou uma mistura entre esses valores) dos alvos. Para garantir que os resultados não serão afetados pelas distintas formas de representar os alvos, basta o usuário utilizar um método de *machine learning* que não seja sensível à forma de representação dos alvos.

A Figura 2.2 apresenta os erros RMSE em função da forma de representação dos alvos e dos métodos de *machine learning* para cada banda predita com base nos satélites Cbers-4, Landsat-8 e Resourcesat-2. O teste t revelou a existência de igualdade (ou diferença) estatística entre alguns RMSE's resultantes dos modelos de predição. Os destaques foram os modelos cujo RMSE foi mínimo bem como os modelos considerados estatisticamente iguais ao mínimo.

A representação dos alvos por meio do valor mínimo (Min) foi a que mais ocorreu os menores RMSE's. Foram 12 ocasiões dentre os 18 RMSE's mínimos referentes a seis bandas preditas com base em três satélites distintos (Figura 2.4). É possível que a representação dos alvos por meio do valor mínimo (Min) tenha sofrido menos com a contaminação dos pixels por reflectâncias de alvos vizinhos devido à resolução espacial das imagens oriundas dos distintos satélites.

A representação Min junto do método de árvore de classificação e regressão com agregação Bootstrap (treebag) apresentou menores RMSE's em oito situações. É característico do método treebag gerar dados adicionais a partir dos dados originais para treinar o modelo, o que diminui o desvio em uma predição e aumenta a sua acurácia (SHI et al., 2019). Esse também pode ter sido o motivo do método treebag ter promovido melhores resultados do que outros métodos de *machine learning* e inteligência artificial nos trabalhos de Carter e Liang (2019).

Figura 2.2. RMSE's dos modelos em função da forma de representação dos alvos e dos métodos de *machine learning*.

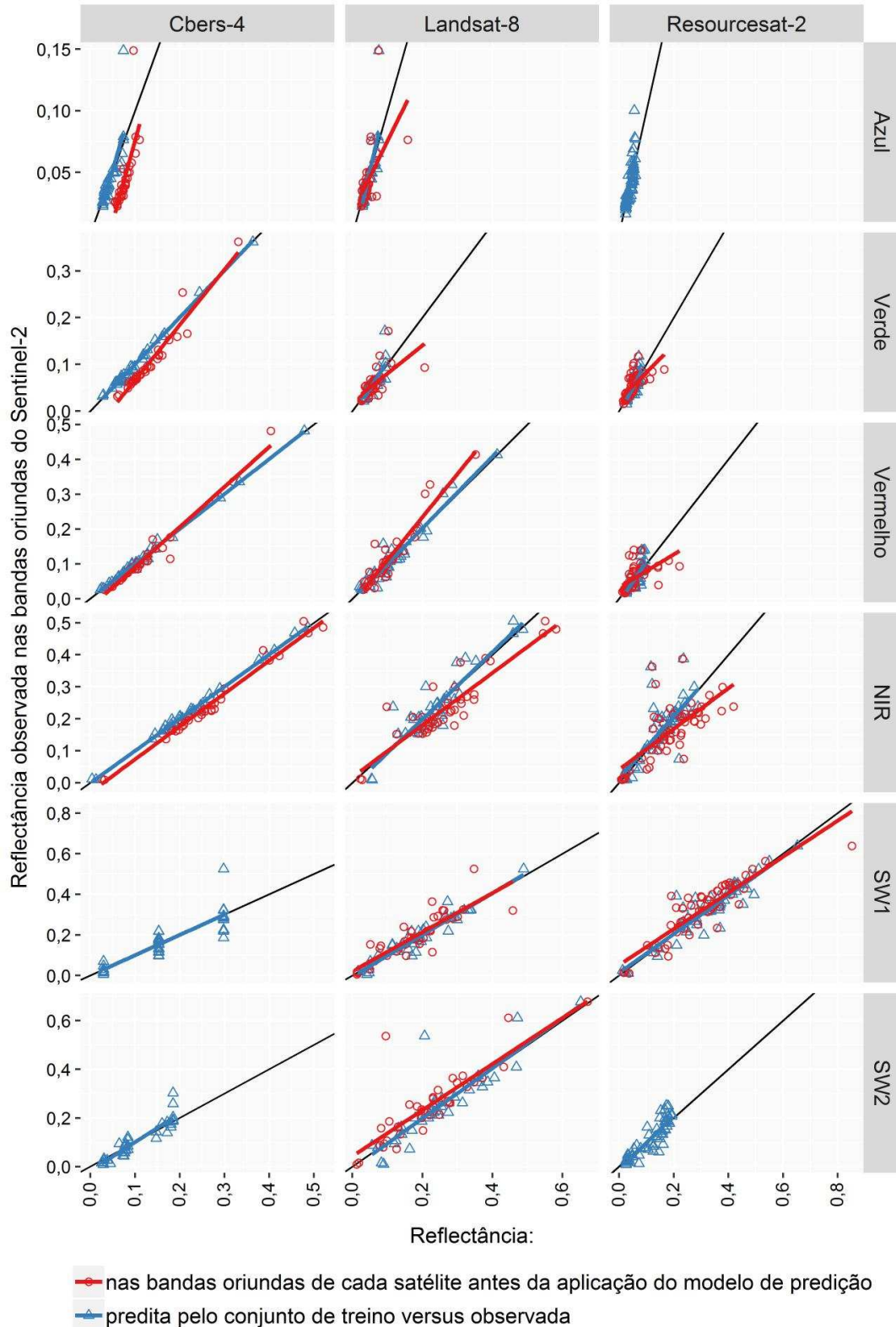


Os métodos de *machine learning* são: regressão linear (lm), máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1), redes neurais artificiais com um neurônio na camada intermediária (nnet1), árvore de classificação e regressão (rpart), árvore de classificação e regressão com agregação Bootstrap (treebag), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

Os valores médios, medianos ou máximos como forma de representar os alvos nas imagens também proporcionaram ajustes cujo RMSE foi mínimo. Porém em uma menor quantidade de vezes do que quando o valor mínimo (forma de representação Min) foi usado. Os métodos de *machine learning* regressão ridge Bayesiana (*Bayesian Ridge Regression* - bridge), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel*- svmLinear1), processo Gaussiano (*Gaussian Process* - gaussprLinear) e a árvore de classificação e regressão (*Classification and Regression Trees* - CART - rpart) também proporcionaram ajustes com RMSE's mínimos.

A Figura 2.3 apresenta as reflectâncias de imagens Sentinel-2 preditas e observadas nas bandas das imagens oriundas dos distintos satélites no conjunto de treinamento e teste dos modelos. Os valores de reflectância preditos com base em imagens Cbers-4 e Resourcesat-2 ficaram menos dispersos do que as reflectâncias originalmente observadas nas bandas oriundas destes satélites. Após a predição da reflectância pelos modelos, os pontos ficaram mais próximos da linha de referência (em que x é igual a y) nas bandas cujas predições utilizaram imagens Cbers-4 e Resourcesat-2 como base dos modelos.

Figura 2.3. Reflectâncias de imagens Sentinel-2 preditas e reflectâncias observadas nas bandas das imagens oriundas dos satélites Cbers-4, Landsat-8, Sentinel-2 e Resourcesat-2 no conjunto de treino dos modelos de predição com RMSE mínimo.



— Linha de referência em que x é igual a y

Dentre todas as predições da banda azul com base em imagens do Cbers-4, 16 modelos apresentaram RMSE estatisticamente iguais ao RMSE mínimo (Figura 2.2). Esse número foi resultado das composições dos modelos pelas formas de representação dos alvos e métodos de mineração. Em relação à banda do verde, vermelho, NIR, SW1 e SW2, a quantidade de modelos considerados com RMSE iguais ao RMSE mínimo foram 17, 20, 27, 25 e 15, respectivamente. Para a predição de uma única banda apenas, pode-se utilizar qualquer um desses modelos listados como melhores.

Contudo, se há o desejo ou necessidade de estimar mais de uma banda utilizando o mesmo modelo, deve-se analisar quais modelos são melhores (menores RMSE's) nas predições de todas as bandas. Os métodos de *machine learning* que estiveram presentes nas listas de melhores modelos para predição das seis bandas, com base em imagens Cbers-4, foram: bridge; gaussprLinear, svmLinear1 e treebag. Nas predições com base em imagens Landsat-8, houve cinco modelos que foram comuns nas listas de melhores modelos em todas as bandas. Semelhante ao ocorrido nas predições com base em imagens Cbers-4, apenas as formas de representação dos alvos Max compuseram os melhores modelos para estimar as seis bandas de imagem. Os melhores modelos para predição das seis bandas também foram compostos pelos métodos de *machine learning* lm, gaussprLinear e rpart.

Quanto às predições com base em imagens oriundas do Resourcesat-2, apenas o modelo criado pelo valor máximo para representar os alvos (Max) em conjunto com o método de *machine learning* lm foi considerado como melhor para predição de todas as seis bandas. Esse modelo foi o que apresentou RMSE estatisticamente igual ao RMSE mínimo resultante das predições de todas as seis bandas.

Em relação às predições utilizando dados de uma data distinta à do treinamento os resultados foram um pouco diferentes. Os modelos que apresentaram os menores RMSE de referência (originado a partir da predição com dados de uma data distinta aos dados de treinamento) não foram os mesmos modelos que apresentaram os menores RMSE's resultantes do conjunto de dados reservado para o teste dos modelos. A Tabela 2.5 apresenta os RMSE's da predição com base no conjunto de dados de avaliação dos modelos cujo erro foi mínimo. A Tabela 2.5 também mostra os RMSE's que esses modelos apresentaram quando o

conjunto de teste dos modelos foi usado para calcular o erro.

Tabela 2.5. Erro em relação a predição de uma data distinta da que foi utilizada para treinar o modelo (modelos com RMSE mínimo na avaliação)

	RMSE de referência • 10 ⁻² (conjunto de avaliação)	RMSE • 10 ⁻² (conjunto de teste)
Cbbers-4		
B.Min.gaussprLinear	1,17	1,55
G.Min.gaussprLinear	1,48	1,87
R.Min.gaussprLinear	2,07	2,76
NIR.Mean.gaussprLinear	2,79	1,42
SW2.Min.treebag	3,62	3,25
SW1.Min.gaussprLinear	4,57	6,28
Resourcesat-2		
B.Mean.treebag	0,88	2,60
G.Min.svmLinear1	1,48	1,92
R.Min.lm	1,93	4,54
SW2.Min.svmLinear1	3,04	5,90
SW1.Max.svmLinear1	4,02	5,43
NIR.Max.treebag	6,01	8,53
Landsat-8		
G.Mean.bridge	1,44	1,88
B.Mean.bridge	1,50	1,99
R.Median.bridge	1,97	3,45
SW2.Mean.bridge	2,37	3,56
NIR.Mean.bridge	2,42	4,83
SW1.Median.bridge	6,11	4,77

Os métodos de *machine learning* são: máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1) e função de custo igual a dois (svmLinear12), árvore de classificação e regressão com agregação Bootstrap (treebag), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

Para as predições com base em imagens do Cbers-4, o método que apresentou o menor valor do RMSE de referência foi o gaussprLinear nas predições de cinco bandas enquanto o método treebag apresentou o menor RMSE de

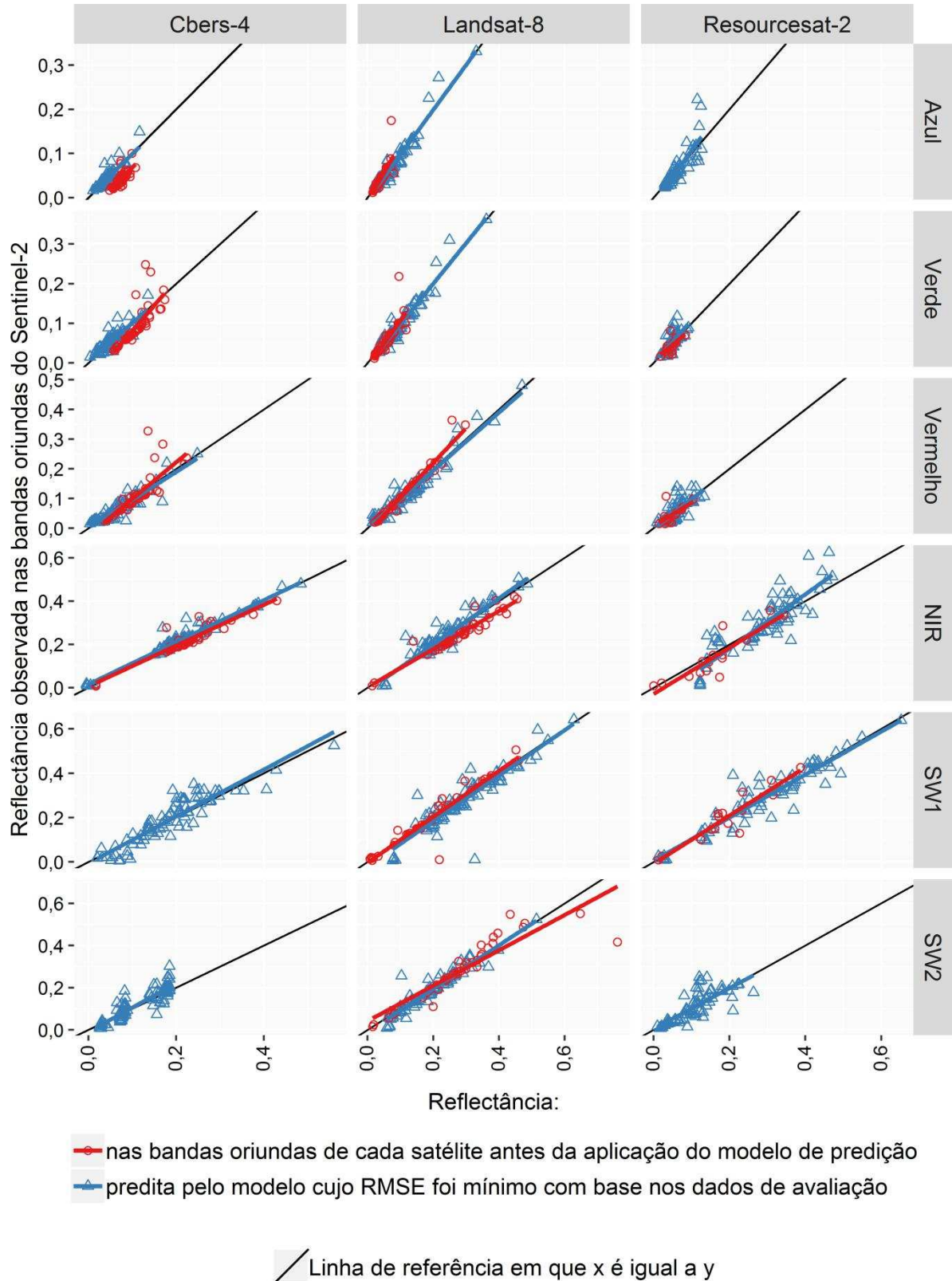
referência apenas na predição de uma banda. Nas predições das bandas B, G, R e SW1, o RMSE de referência foi menor do que o RMSE resultante das predições com o conjunto de teste dos modelos.

Nas predições com base em imagens do Resourcesat-2, três métodos apresentaram menor valor do RMSE de referência nas predições de seis bandas. Desses três métodos, o svmLinear1 apresentou o menor RMSE para três bandas que foram elas: o verde (G) e o infravermelho de ondas curtas um e dois (SW1 e SW2, respectivamente). Nas predições de todas as bandas, o RMSE de referência foi menor do que o RMSE calculado por com os dados de teste dos modelos.

O método bridge apresentou os menores RMSE's de referência nas predições de todas as bandas com base em imagens Landsat-8. O RMSE de referência foi menor do que o RMSE do conjunto de teste nas predições de 5 bandas. Esses resultados comuns às predições tanto com base no Cbers-4, quanto no Resourcesat-2 e no Landsat-8 nos revela que é possível estimar a reflectância das imagens do Sentinel-2 para distintas datas com erros até menores do que os erros de desempenho dos modelos (calculados com os dados reservados para teste dos modelos utilizando a técnica de validação cruzada).

A Figura 2.4 apresenta as reflectâncias de imagens Sentinel-2 preditas e as reflectâncias observadas nas bandas das imagens oriundas dos satélites Cbers-4, Landsat-8, Sentinel-2 e Resourcesat-2 no conjunto de avaliação dos modelos. Os valores de reflectância nas bandas azul, verde, SW1 e SW2 preditos com base em imagens Landsat-8 ficaram menos dispersos do que as reflectâncias originalmente observadas nas bandas oriundas deste satélite. O mesmo ocorreu com as reflectâncias nas bandas do vermelho, NIR e SW1 preditas com base em imagens Resourcesat-2. Após a predição da reflectância pelos modelos, os pontos ficaram mais próximos da linha de referência (em que x é igual a y) nas bandas do azul, verde e vermelho cujas predições utilizaram imagens Cbers-4 e nas bandas do vermelho, verde e NIR cujas predições ocorreram com base em imagens Resourcesat-2.

Figura 2.4. Reflectâncias de imagens Sentinel-2 preditas e reflectâncias observadas nas bandas das imagens oriundas dos satélites Cbers-4, Landsat-8, Sentinel-2 e Resourcesat-2 no conjunto de avaliação dos modelos de predição com erro mínimo.

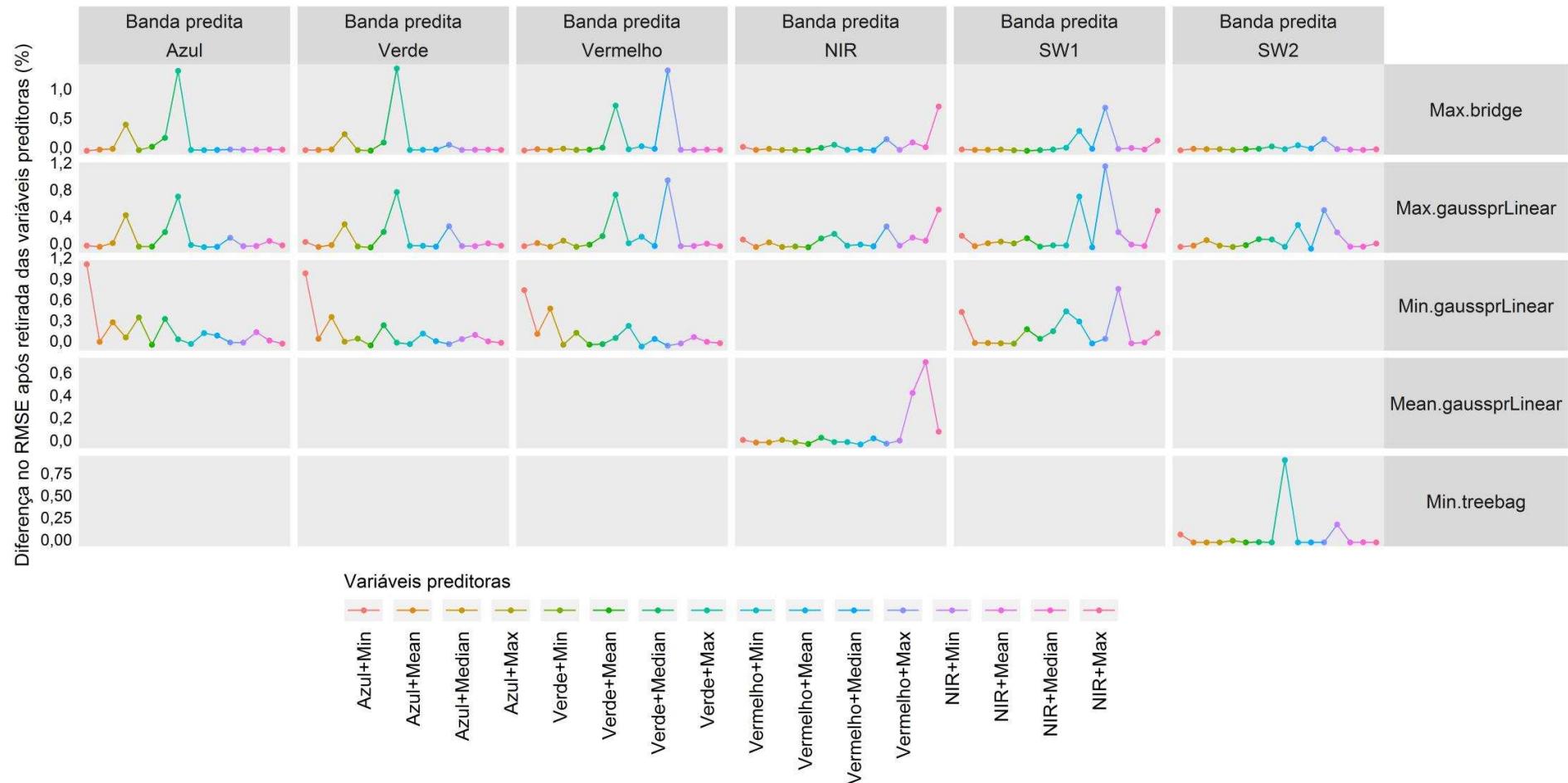


A Figura 2.5 apresenta a importância das variáveis preditoras nos modelos de predição da reflectância de imagens Sentinel-2 com base em imagens oriundas do Cbers-4. Os modelos listados foram aqueles com menores RMSE's na predição das seis bandas simultaneamente. Além desses, a importância das variáveis preditoras também foi verificada nos modelos com RMSE de referência mínimo quando avaliados pelo conjunto de dados reservados para avaliação. De modo semelhante, a Figura 2.6 e a Figura 2.7 também mostram as variáveis importantes nas predições de imagens Sentinel-2. Porém, a predição ocorreu com base em imagens oriundas dos satélites Resourcesat-2 e Landsat-8, respectivamente.

O ajuste dos modelos com uma variável de entrada a menos gerou um RMSE diferente do RMSE dos modelos criados com todas as variáveis de entrada e calculado com base no conjunto de avaliação dos modelos (RMSE de referência). Nas predições com base em imagens oriundas do Cbers-4, essa diferença foi de até 1,4% (Figura 2.5). Nas predições com base em imagens Resourcesat-2, a diferença entre os RMSE's foi de no máximo 7,68% (Figura 2.6). Já nas predições da reflectância nas bandas das imagens Sentinel-2 utilizando como entrada as bandas das imagens Landsat-8, essa diferença foi de no máximo 16,73% (Figura 2.7). As predições com base em imagens Landsat-8 foram mais sensíveis às variáveis de entrada, pois, a ausência delas gerou RMSE's mais diferentes do RMSE de referência do que quando a predição usou imagens oriundas dos demais satélites.

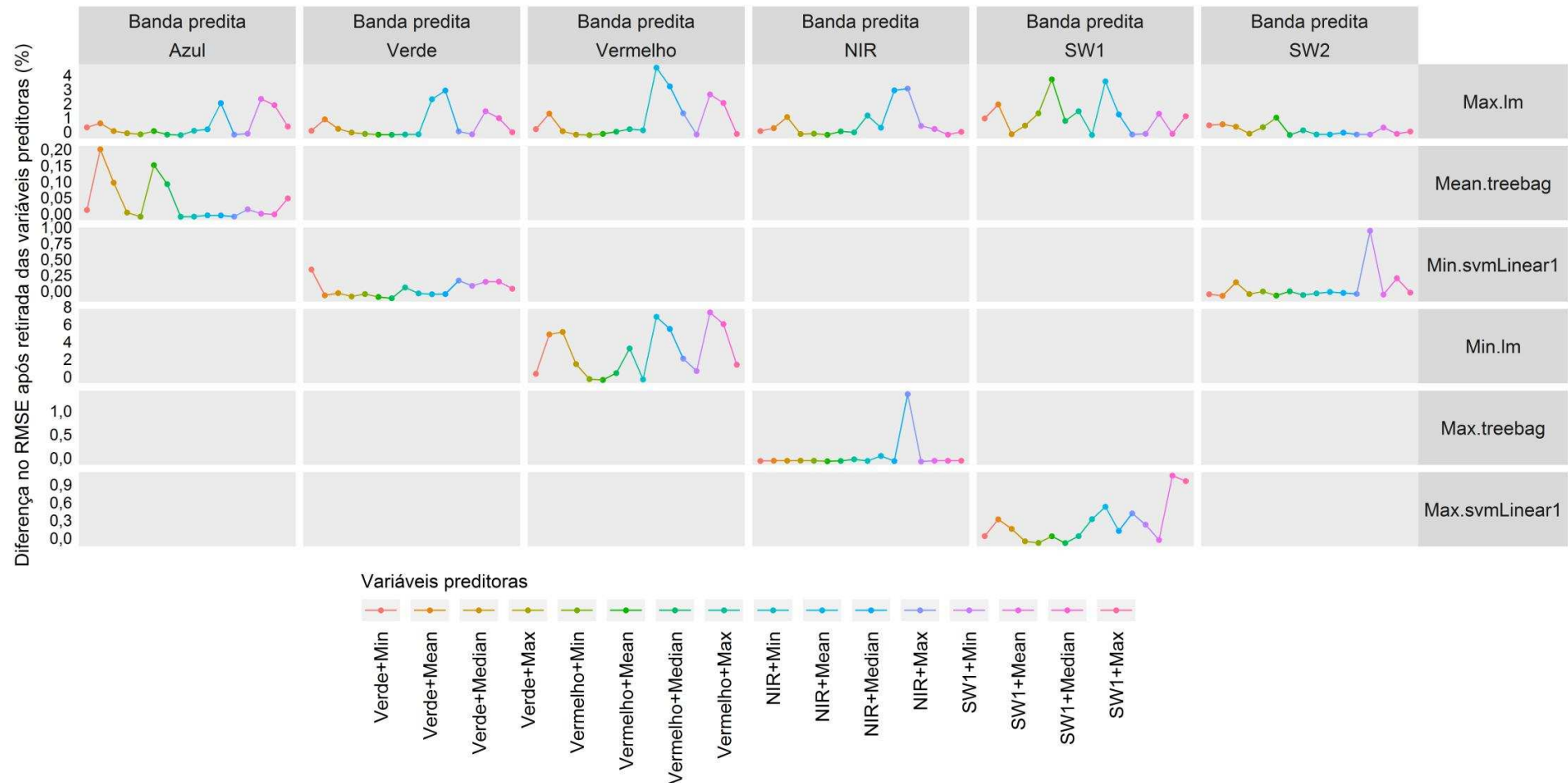
Em alguns dos casos ilustrados, a retirada de variáveis preditoras dos modelos gerou pouca diferença em relação ao RMSE de referência. A pequena diferença entre os RMSE's demonstrou que uma vez que uma variável preditora foi retirada do treinamento do modelo, outras variáveis conseguiram realizar a predição tão bem quanto à predição com todas as variáveis preditoras. Porém, a variável mais importante (ou a que mais contribuiu na predição) foi aquela cuja ausência no ajuste do modelo causou a maior diferença no RMSE. Essas diferenças podem ser vistas na Figura 2.5 e na Figura 2.7 nos gráficos em que a curva aparece constante no nível de 0%. A predição da banda SW2 pelo modelo Max.bridge com base em imagens Cbers-4 e a predição da banda SW1 pelo modelo Median.bridge com base em imagens Landsat-8 são exemplos de casos em que a variável ausente no treinamento não causou grande diferença no RMSE.

Figura 2.5. Importância das variáveis preditoras nos modelos com base em imagens oriundas do Cbers-4.



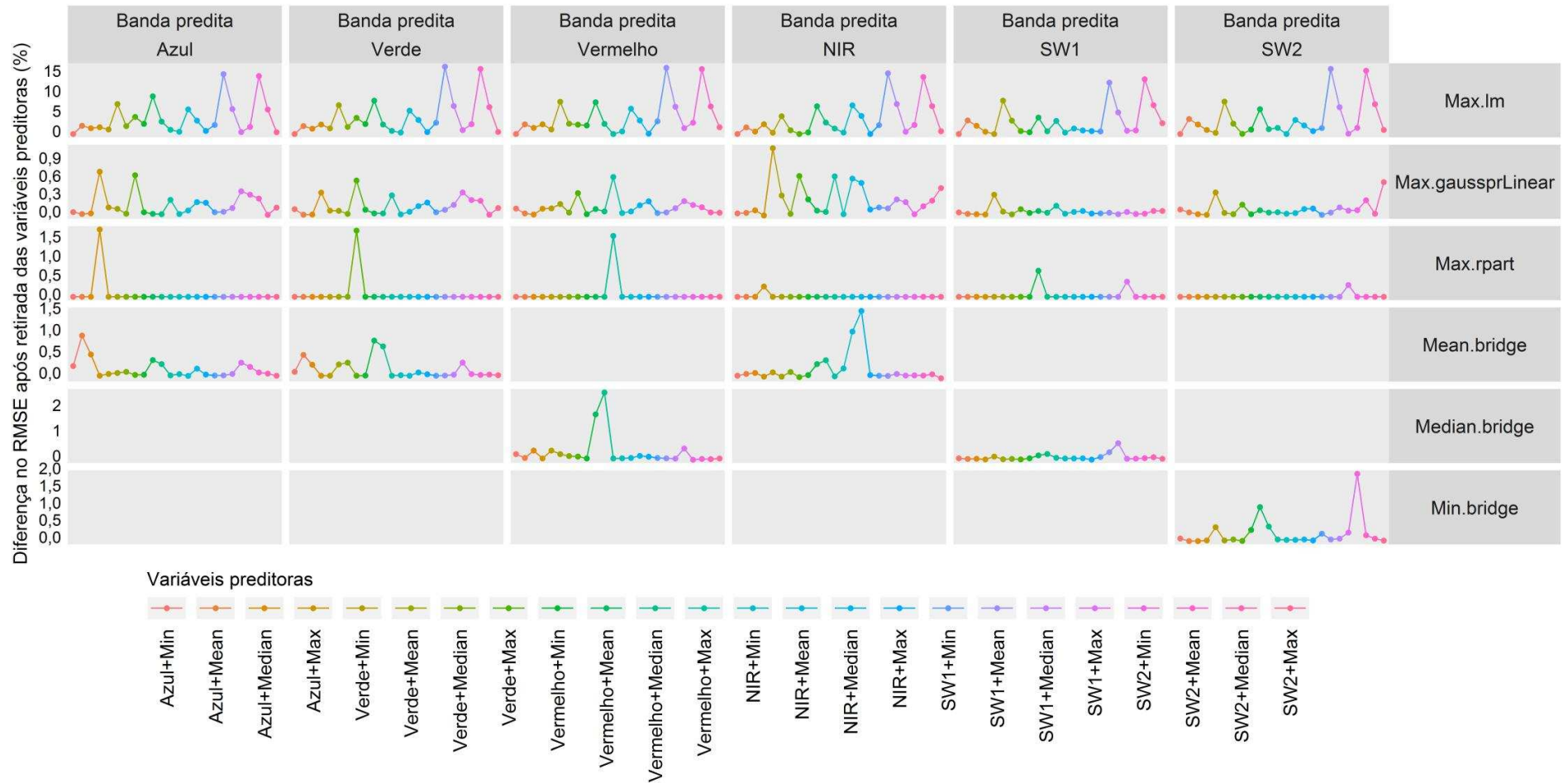
Os métodos de *machine learning* são: árvore de classificação e regressão com agregação Bootstrap (treebag), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente. O símbolo + indica que a variável preditora corresponde à reflectância em cada banda de acordo com as formas de representação dos alvos.

Figura 2.6. Importância das variáveis preditoras nos modelos com base em imagens oriundas do Resourcesat-2.



Os métodos de *machine learning* são: regressão linear (lm), máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1) e árvore de classificação e regressão com agregação Bootstrap (treebag). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente. O símbolo + indica que a variável preditora corresponde à reflectância em cada banda de acordo com as formas de representação dos alvos.

Figura 2.7. Importância das variáveis preditoras nos modelos com base em imagens oriundas do Landsat-8.



Os métodos de *machine learning* são: regressão linear (lm), árvore de classificação e regressão (rpart), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente. O símbolo + indica que a variável preditora corresponde à reflectância em cada banda de acordo com as formas de representação dos alvos.

Em alguns casos, foi observado que apenas uma variável se sobressaiu às demais quanto à diferença no RMSE causada por sua ausência. Nas previsões com base em imagens Cbers-4, isso pode ser visto na previsão da banda NIR pelo modelo Max.bridge. Nessa situação, a variável preditora mais importante foi a reflectância na banda do verde cuja representação dos alvos na imagem ocorreu pelo valor máximo da reflectância nos alvos (Verde+Max). Na Figura 2.5, a diferença no RMSE causada pela ausência dessa variável fez o ponto se sobressair à curva no nível de 0%.

Nas previsões das bandas do azul, verde e vermelho pelo modelo Max.rpart com base em imagens Landsat-8 (Figura 2.7) também houve apenas uma variável que se destacou das demais quanto a sua importância. As variáveis que mais contribuíram nas previsões das bandas do azul, verde e vermelho do Sentinel-2 foram, respectivamente, as bandas do azul, verde e vermelho das imagens Landsat-8. Em todas essas bandas, a reflectância dos alvos nas bandas correspondeu ao valor máximo (Azul+Max, Verde+Max, Vermelho+Max).

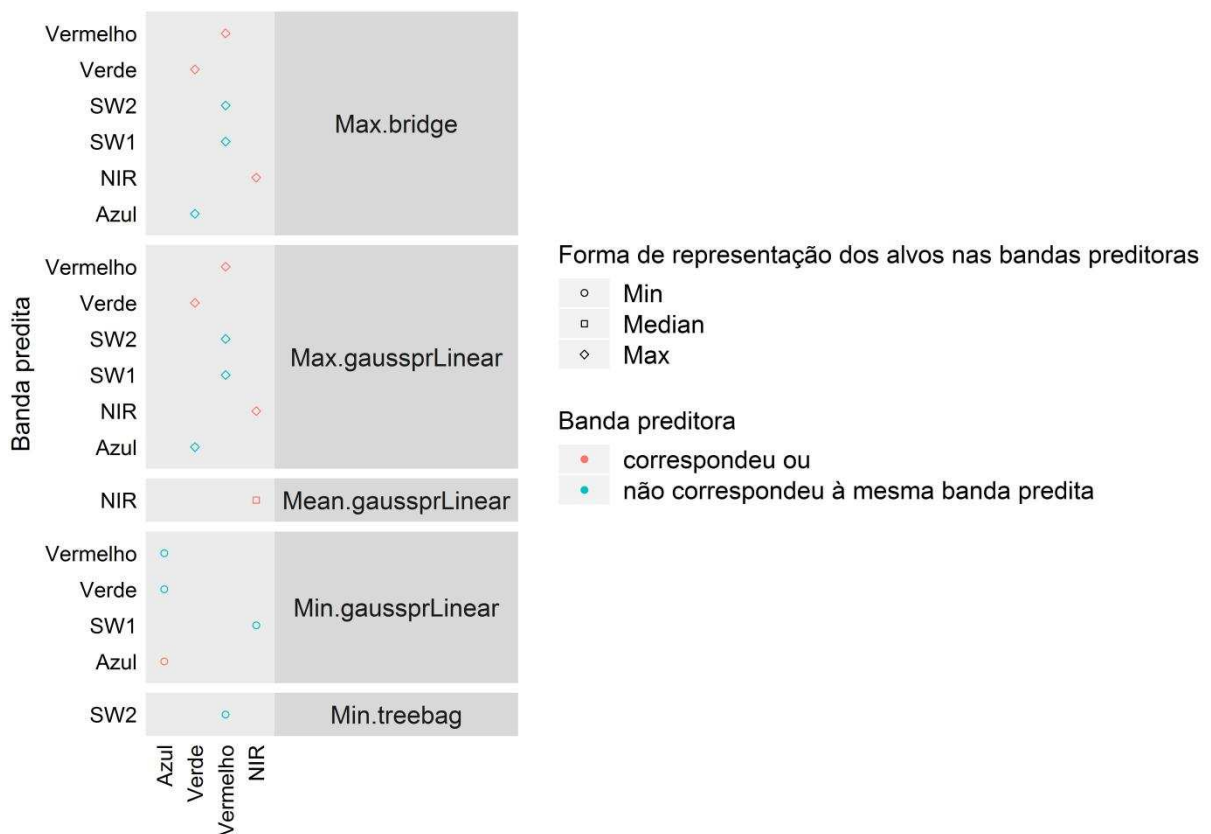
Na Figura 2.5, na Figura 2.6 e na Figura 2.7, é possível visualizar, ainda, os casos em que a diferença no RMSE causada pela ausência de uma segunda, terceira ou quarta variável de maior importância se sobressaiu às demais. Isso pode ser observado principalmente em relação às previsões com base em imagens Resourcesat-2 (Figura 2.6). Nesse caso, nove dentre os 12 modelos que predisseram as bandas do Sentinel-2 com base em imagens Resourcesat-2 apresentaram mais de uma variável importante.

A Figura 2.8 destaca as variáveis mais importantes nas previsões das bandas de imagens Sentinel-2 com base em imagens oriundas do Cbers-4. Em sete situações, do total de 18 apresentadas na Figura 2.5, a banda do vermelho foi a mais importante na previsão das bandas do vermelho, SW1 e SW2 das imagens Sentinel-2. Isto é, a banda do vermelho oriunda do Cbers-4 foi importante na previsão de outras bandas além do vermelho de imagens Sentinel-2. Nesses casos, a banda do vermelho oriunda do Cbers-4 foi importante, especialmente, nas previsões das bandas do SW1 e SW2 de imagens Sentinel-2. Esse resultado ocorreu devido à contribuição que a banda do vermelho forneceu e também devido ao fato das bandas do SW1 e SW2 não terem composto o banco de dados de entrada dos modelos, pois elas não foram capturadas pelo sensor MUX a bordo do Cbers-4. Logo, as bandas SW1 e SW2 não poderiam contribuir com a modelagem pois elas

foram inexistentes no banco de dados.

Além do vermelho, as bandas do azul, verde e NIR capturadas pelo sensor a bordo do Cbers-4 também foram importantes nas predições de outras bandas de imagens Sentinel-2 além do azul, verde e NIR, respectivamente. A banda do azul oriunda do Cbers-r4 foi importante na predição da banda do verde e do vermelho no modelo Min.gaussprLinear. A banda do verde oriunda do Cbers-4 foi a mais importante na predição da banda do azul nos modelos Max.bridge e Max.gaussprLinear. A banda do NIR oriunda do Cbers-4 foi a mais importante na predição da banda do SW1 pelo modelo Min.gaussprLinear.

Figura 2.8. Variáveis mais importantes nas predições das bandas de imagens Sentinel-2 em função de imagens oriundas do Cbers-4 nos modelos que apresentaram os menores RMSE's no conjunto de teste e avaliação dos modelos.

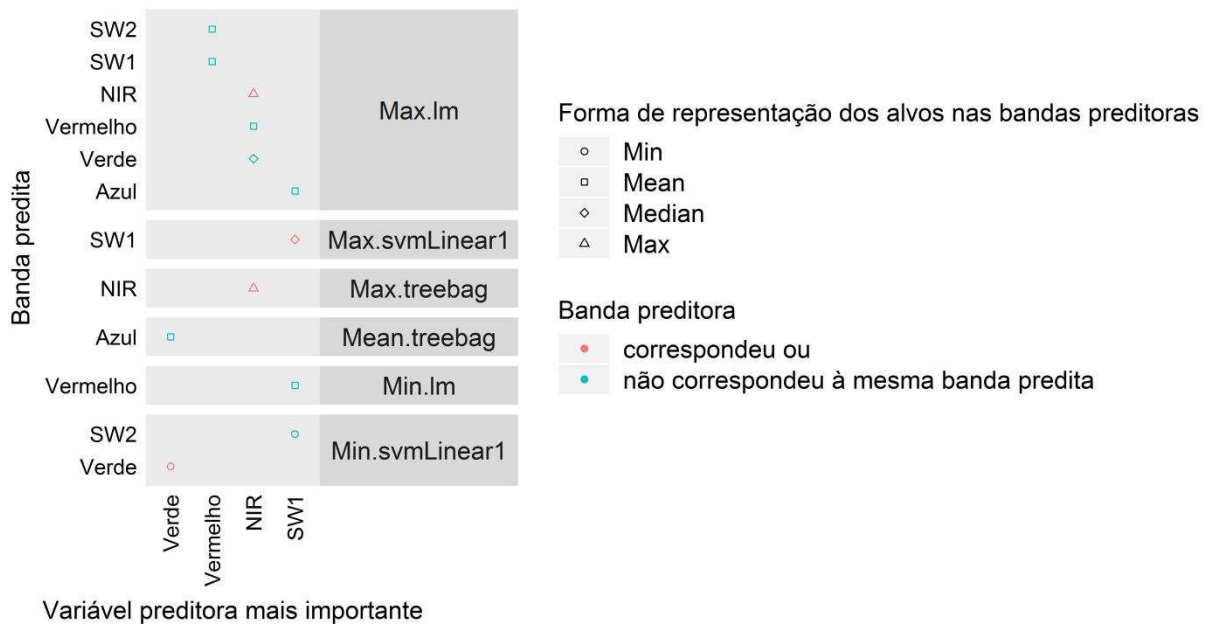


Os métodos de *machine learning* são: árvore de classificação e regressão com agregação Bootstrap (treebag), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

A Figura 2.9 ilustra as variáveis mais importantes nas predições das bandas

de imagens Sentinel-2 com base em imagens oriundas do Resourcesat-2. As bandas do NIR e do SW1 oriundas do Resourcesat-2 foram as que mais contribuíram nas predições das bandas de imagens Sentinel-2. Cada uma compôs a variável mais importante na predição em quatro casos, dentre o total de 12 ilustradas na Figura 2.6 e na Figura 2.9.

Figura 2.9. Variáveis mais importantes nas predições das bandas de imagens Sentinel-2 em função de imagens oriundas do Resourcesat-2.



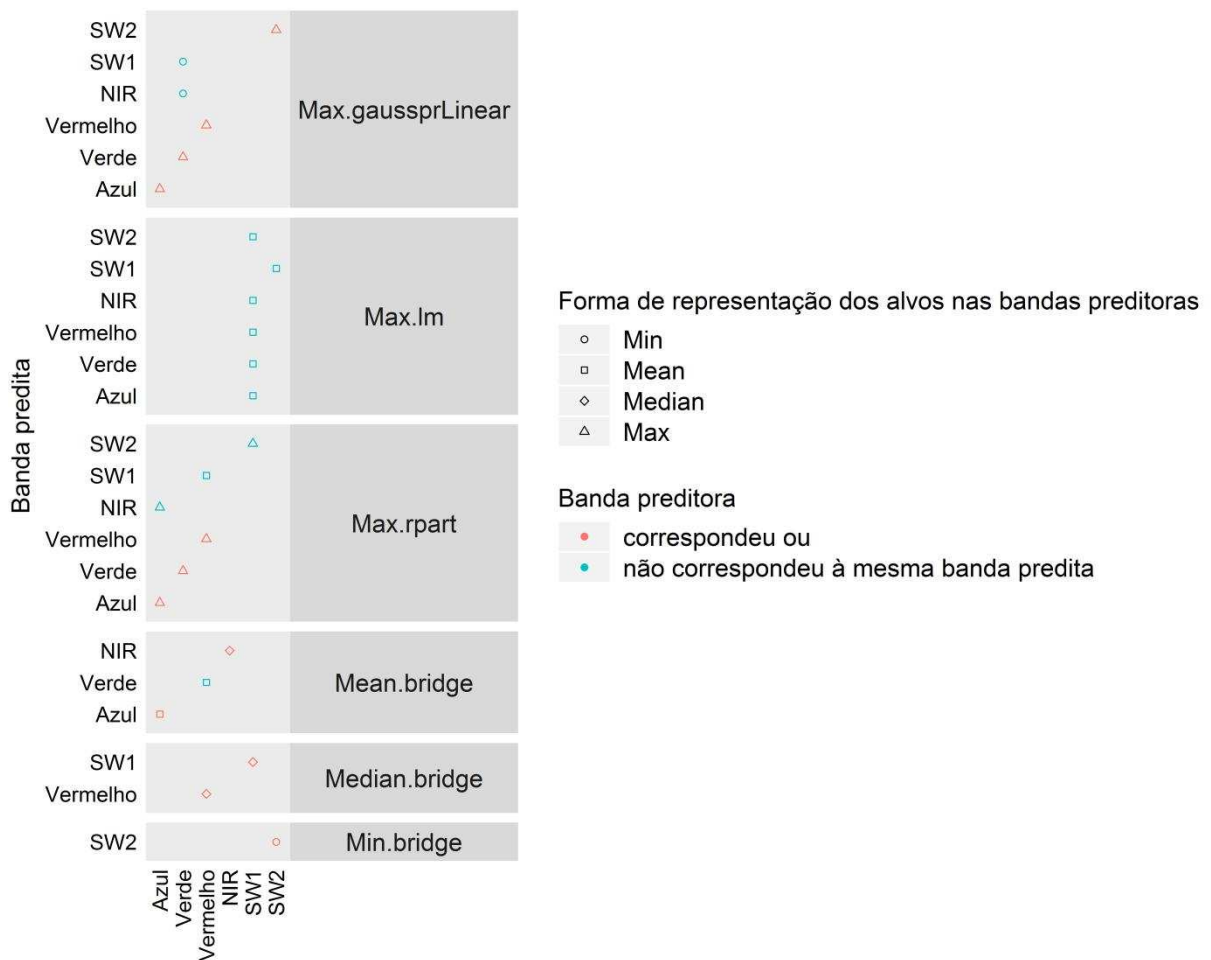
Os métodos de *machine learning* são: regressão linear (lm), máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1) e árvore de classificação e regressão com agregação Bootstrap (treebag). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

Semelhante ao ocorrido nas predições com base em imagens Cbers-4, as bandas do verde, do NIR e do SW1 oriundas do Resourcesat-2 foram importantes nas predições de outras bandas de imagens Sentinel-2 além do verde, do NIR e do SW1. As variáveis mais importantes na predição da banda do azul foram as bandas do verde e do SW1. A banda verde e a banda SW1 foram mais importantes quando, respectivamente, os modelos Mean.treebag e Max.lm foram usados para estimar a banda do azul. A banda SW1 também foi a mais importante na predição da banda SW2 quando, para isto, o modelo Min.svmLinear1 foi usado. Já quando a predição da banda SW2 foi realizada pelo modelo Max.lm, a variável mais importante foi composta pela banda do vermelho oriunda do Resourcesat-2. As bandas do azul e

do SW2 não compuseram o conjunto de dados de entrada dos modelos, pois não foram capturadas pelo sensor LISS-III do Resourcesat-2.

A Figura 2.10 destaca as variáveis mais importantes nas predições das bandas de imagens Sentinel-2 com base em imagens capturadas pelo sensor a bordo do Landsat-8. Em 7 casos, do total de 24 apresentados na Figura 2.10, a banda do SW1 foi a mais importante na predição das seis bandas das imagens Sentinel-2. Em apenas um caso, a banda do SW1 oriunda do Landsat-8 predisse a banda do SW1 das imagens Sentinel-2. Isso ocorreu quando o modelo Median.bridge foi usado na predição.

Figura 2.10. Variáveis mais importantes nas predições das bandas de imagens Sentinel-2 em função de imagens oriundas do Landsat-8.



Os métodos de *machine learning* são: regressão linear (lm), árvore de classificação e regressão (rpart), regressão ridge Bayesiana (bridge) e processo Gaussiano (gaussprLinear). Max, Mean, Median e Min são as formas de representar os alvos por meio dos valores máximo, médio, mediano e mínimo, respectivamente. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

As bandas do azul, verde, vermelho e SW2 originadas no Landsat-8 também foram importantes nas predições de outras bandas de imagens Sentinel-2 além do azul, verde, vermelho e SW2, respectivamente. Esse fato de uma banda de um satélite ser a mais importante na predição de uma banda distinta do outro satélite foi comum a várias predições com base no Cbers-4, no Resourcesat-2 e no Landsat-8, como mencionado. Ou seja, houve casos em que a banda que compôs a variável preditora mais importante não correspondeu a mesma banda predita nas predições com base nos três satélites. Então, para estimar uma banda de um satélite pode-se utilizar outras bandas de outros satélites.

2.4 CONCLUSÃO

Os métodos de regressão linear (lm), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel- svmLinear1*), redes neurais artificiais (nnet1), árvore de classificação e regressão (*Classification and Regression Trees - CART*), CART com agregação Bootstrap (*Bootstrap Aggregating - bagged CART*), regressão ridge Bayesiana (*Bayesian Ridge Regression*) e processo Gaussiano (*Gaussian Process - gaussprLinear*) mostraram-se eficazes para estimar os valores de reflectância de imagens Sentinel-2 com base em imagens oriundas do Cbers-4, do Landsat-8 e do Resourcesat-2.

O método de redes neurais artificiais foi menos sensível à forma de representação dos alvos do que os demais métodos para predição de imagens orbitais.

Os métodos que apresentaram menores RMSE's na predição da reflectância de imagens orbitais em uma data diferente daquela a qual os dados serviram para treinar os modelos foram: processo Gaussiano (*Gaussian Process - gaussprLinear*) e árvore de classificação e regressão com agregação Bootstrap (treebag) para as predições com base em imagens Cbers-4; regressão linear (lm), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel- svmLinear1*) e treebag nas predições com base em imagens Resourcesat-2; e, regressão ridge Bayesiana (*Bayesian Ridge Regression - bridge*) nas predições com base em imagens Landsat-8.

Houve situações em que a banda espectral utilizada na variável preditora

mais importante dos modelos foi diferente da banda espectral predita.

2.5 REFERÊNCIAS

AHMAD, F.; GOPARAJU, L.; QAYUM, A.. Natural resource mapping using Landsat-8 and Lidar towards identifying digital elevation, digital surface and canopy height models. **International Journal of Environmental Sciences and Natural Resources**, v. 2, 2017.

BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G. Configuração de redes neurais artificiais para estimação do volume de árvores. **Revista Ciência da Madeira**, v. 5, n. 1, 2014.

CAI, Y. et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. **Agricultural and Forest Meteorology**, v. 274, 2019.

CARTER, C.; LIANG, S.. Evaluation of ten machine learning methods for estimating terrestrial evapotranspiration from remote sensing. **International Journal of Applied Earth Observation and Geoinformation**, v. 78, 2019.

CEPSRM (Rio Grande do Sul). Universidade Federal do Rio Grande do Sul. **Página Dinâmica para Aprendizado do Sensoriamento Remoto: Sensores e Plataformas Orbitais**. Disponível em: <<http://www.ufrgs.br/engcart/PDASR/sensores.html>>. Acesso em: 06 fev. 2017.

DU, Xin et al. Detecting advanced stages of winter wheat yellow rust and aphid infection using RapidEye data in North China Plain. **GIScience and Remote Sensing**, 2019.

EPIPHANIO, J. C. N.. CBERS-4: estado atual e futuro. **XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Rio Grande do Norte**, 2009.

FISHER, R. P.; HOBGEN, S. E.; HALEBEREK, K.; SULA, N.; MANDAYA, I.. Free satellite imagery and digital elevation model analyses enabling natural resource management in the developing world: case studies from Eastern Indonesia. **Singapore Journal of Tropical Geography**, v. 39, n. 1, 2018.

GURU, B.; SESHAN, K.; BERA, S.. Frequency ratio model for groundwater potential mapping and its sustainable management in cold desert, India. **Journal of King Saud University-Science**, v. 29, n. 3, 2017.

HALLETT, S. H.; SAKRABANI, R.; KEAY, C. A.; HANNAM, J. A.. Developments in land information systems: examples demonstrating land resource management capabilities and options. **Soil use and management**, v. 33, n. 4, 2017.

KEERTHI, V., KUMAR, A. S. At-sensor solar exo-atmospheric irradiance, Rayleigh optical thickness and spectral parameters of RS-2

sensors.NRSC/SDAPSA/DQEPQCD/RS-2/TN-July2011. 2011.

LEAL, F. A.; MIGUEL, E. P.; MATRICARDI, E. A. T.; PEREIRA, R. S. Redes neurais artificiais na estimativa de volume em um plantio de eucalipto em função de fotografias hemisféricas e número de árvores. **Revista Brasileira de Biometria**, v. 33, n. 2, 2015.

MCCARTHY M. J.; COLNA K. E.; EL-MEZAYEN M. M.; LAUREANO-ROSARIO A. E.; MÉNDEZ-LÁZARO P.; OTIS D. B.; TORO-FARMER G.; VEGA-RODRIGUEZ M.; MULLER-KARGER F. E.. Satellite remote sensing for coastal management: A review of successful applications. **Environmental management**, v. 60, n. 2, 2017.

MOSE, V. N.; WESTERN, D.; TYRRELL, P.. Application of open source tools for biodiversity conservation and natural resource management in East Africa. **Ecological informatics**, v. 47, 2018.

NIGAM, R.; VYAS, S. S.; BHATTACHARYA, B. K.; OZA, M. P.; MANJUNATH, K. R. Retrieval of regional LAI over agricultural land from an Indian geostationary satellite and its application for crop yield estimation. **Journal of Spatial Science**, 2016.

OLIVEIRA, A. C. S.; SOUZA, A. A.; LACERDA, W. S.; GONÇALVES, L. R. Aplicação de redes neurais artificiais na previsão da produção de álcool. **Ciênc.agrotec**. v. 34, n. 2, 2010.

PETERSEN, L..Real-time prediction of crop yields from MODIS relative vegetation health: a continent-wide analysis of Africa. **Remote Sensing**, v. 10, n. 11, 2018.

PINTO, C.; PONZONI, F.; CASTRO, R.; LEIGH, L.; MISHRA, N.; AARON, D.; HELDER, D.. First in-flight radiometric calibration of MUX and WFI on-board CBERS-4. **Remote Sensing**, v. 8, n. 5, 2016.

TEAM, R. Core et al. R: **A language and environment for statistical computing**.R Foundation for Statistical Computing, Vienna, 2018.

SANHOUSE-GARCIA, A. J.; BUSTOS-TERRONES, Y.; RANGEL-PERAZA, J. G.; QUEVEDO-CASTRO, A.; PACHECO, C. Multi-temporal analysis for land use and land cover changes in an agricultural region using open source tools. **Remote Sensing Applications: Society and Environment**, v. 8, 2017.

SANHOUSE-GARCÍA, A. J.; RANGEL-PERAZA, J. G.; BUSTOS-TERRONES, Y.; GARCÍA-FERRER, A.; MESAS-CARRASCOSA, F. J. Land use mapping from CBERS-4-2 images with open source tools by applying different classification algorithms. **Physics and Chemistry of the Earth, Parts A/B/C**, v. 91, 2016.

SHI, H.; HWANG, K. S.; LI, X.; CHEN, J. A learning approach to image-based visual servoing with a bagging method of velocity calculations. **Information Sciences**, v. 481, 2019.

SHIMRAH, T.; SARMA, K.; VARGA, O. G.; SZILARD, S.; SINGH, S. K.. Quantitative assessment of landscape transformation using earth observation datasets in Shirui

Hills of Manipur, India. **Remote Sensing Applications: Society and Environment**, 2019.

SHIU, Y. S.; CHUANG, Y. C..Yield estimation of paddy rice based on satellite imagery: comparison of global and local regression models. **Remote Sensing**, v. 11, n. 2, 2019.

SONG, R.; CHENG, T.; YAO, X.; TIAN, Y.; ZHU, Y.; CAO, W. Evaluation of Landsat-8 time series image stacks for predicting yield and yield components of winter wheat. In: **Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International**. IEEE, 2016.

YUE, J.; YANG, G.; FENG, H. Comparative of remote sensing estimation models of winter wheat biomass based on random forest algorithm. **Transactions of the Chinese Society of Agricultural Engineering**, v. 32, n. 18, 2016.

3. Estimativa da produtividade por meio de perfis espectrais do cafeeiro e métodos de *machine learning*

RESUMO

A produtividade da lavoura influencia algumas atividades agrícolas e sua estimativa pode ser útil aos produtores. Existem alguns modelos que tentam estimar a produtividade, porém, a quantidade de variáveis necessárias e a dificuldade em mensurá-las são um problema. Como a resposta espectral da planta é resultado do seu vigor vegetativo, que por sua vez, relaciona-se à produtividade agrícola (PICOLI et al., 2009), objetivou-se neste trabalho estimar a produtividade do café utilizando informações espectrais da cultura e métodos de *machine learning*. As informações espectrais foram obtidas a partir de imagens orbitais e corresponderam a seis índices de vegetação e reflectância em seis bandas espectrais das imagens. A estimativa da produtividade ocorreu por meio de seis métodos de *machine learning*. Foram eles: regressão linear, máquinas de vetores de suporte com Kernel linear, redes neurais artificiais, regressão ridge Bayesiana, processo Gaussiano e floresta aleatória. Os modelos foram implantados em linguagem R no programa computacional R Versão 3.5.1 (R Team, 2018). A raiz do erro quadrático médio (*root mean square error* - RMSE) e o erro médio absoluto (*mean absolute error* - MAE) foram usados para avaliar a acurácia dos modelos. O RMSE e o MAE serviram de entrada para o teste de Scott-Knott que agrupou os modelos semelhantes. A estimativa da produtividade com o uso de apenas cinco variáveis apresentou erros semelhantes, a 5% de significância, aos erros resultantes dos modelos com 10 e com 14 variáveis que incluíram informações espectrais, topográficas e agronômicas. O erro RMSE mínimo apresentado pelos modelos correspondeu a uma diferença de 11% entre o valor estimado e o valor real da produtividade do café do talhão Pasto Novo 1 no ano de 2017. O erro MAE mínimo correspondeu a uma diferença de 1,7% entre o valor estimado e o valor observado da produtividade do talhão Açude 3 no ano de 2018. Os erros resultantes dos modelos de *machine learning* foram menores do que os erros do método fundamentado no Índice Fenológico de Produção (IFP).

Palavras-chave: redes neurais artificiais, máquinas de vetores de suporte, regressão linear, ciência dos dados.

ABSTRACT

The crop yield influences some agricultural activities and its estimation can be useful to the farmers. There are some models that try to estimate the yield, but the amount of variables needed and difficulty in measuring them is a problem. As the spectral response of the plant is a result of its vegetative vigor, which in turn is related to the agricultural yield (PICOLI et al., 2009), the objective of this work was to estimate coffee yield using spectral information from the crop and methods of machine learning. The spectral information was obtained from orbital images and corresponded to six indices of vegetation and reflectance in six spectral bands of the images. The estimate of yield occurred through six methods of machine learning. They were: linear regression, support vector machines with linear kernel, artificial neural networks, Bayesian ridge regression, Gaussian process and random forest. The models were implanted in R language in the computer program R Version 3.5.1 (R Team, 2018). The root mean square error (RMSE) and the mean absolute error (MAE) were used to evaluate the accuracy of the models. The RMSE and MAE served as input to the Scott-Knott test that grouped the similar models. Productivity estimates using only five variables showed similar errors, at 5% of significance, to errors resulting from models with 10 and 14 variables that included spectral, topographic and agronomic information. The minimum RMSE error presented by the models corresponded to a difference of 11% between the estimated value and the real productivity value of Pasto Novo 1 field coffee in 2017. The minimum MAE error corresponded to a difference of 1.7% between the estimated value and the observed productivity value of the Açude 3 field in 2018. The errors resulting from the machine learning models were smaller than the errors of the method based on the Phenological Production Index (IFP).

Key words: artificial neural networks, support vector machines, linear regression, data science.

3.1 INTRODUÇÃO

O café é utilizado como ingrediente de bebidas e comidas em todo o mundo devido ao seu sabor e efeito estimulante. Assim como outros produtos agrícolas, substâncias presentes nos grãos de café também compõem cosméticos e medicamentos. A demanda dos vários mercados consumidores de café é suprida por produtores de mais de 70 países, sendo que destes, apenas o Brasil, o Vietnã e a Indonésia juntos produzem mais de 50% de todo o café produzido no mundo (FAO, 2015).

A produtividade da lavoura influencia algumas atividades agrícolas. O dimensionamento da infraestrutura, contratação de mão-de-obra e verificação de recursos materiais e financeiros são exemplos de tarefas que dependem da produtividade da lavoura para serem realizadas de maneira adequada. Sendo o preço função da demanda e da oferta do produto, a produtividade pode, mesmo que de maneira indireta, influenciar a precificação do produto. Dessa forma, a estimativa da produtividade pode ser útil aos produtores.

A produtividade agrícola é função de muitas variáveis ambientais e de práticas agronômicas na lavoura. Logo, é natural que alguns modelos utilizem informações edafoclimáticas para tentar estimar a produtividade. Um exemplo de modelagem matemática que utiliza essas informações são os modelos agrometeorológicos (GOMES et al., 2014; NUNES et al., 2010). Os modelos agrometeorológicos-espectrais incluem na modelagem informações de índices de vegetação como o Índice de Vegetação da Diferença Normalizada (*Normalized Difference Vegetation Index* – NDVI). Afinal, a resposta espectral da planta é resultado do seu vigor vegetativo, que por sua vez, relaciona-se à produtividade agrícola (PICOLI et al., 2009).

Fahl et al. (2005), visando o desenvolvimento de um modelo para estimar a produtividade independente da cultivar e da densidade de plantio, da idade da cultura e das condições edafoclimáticas, avaliaram características fenológicas determinantes do crescimento e da produção do cafeeiro. Os estudos levaram à obtenção do índice fenológico de produção (IFP) (FAHL; CARELLI, 2007). A comparação entre a produção estimada e a real apresentou uma correlação de 0,97 e uma margem de erro de 7%, enquanto a comparação entre as estimativas efetuadas visualmente, por técnicos especializados na cultura, e as reais, a margem

de erro foi de 9% (FAHL; CARELLI, 2007). A conclusão do estudo foi que a utilização do modelo desenvolvido com base no IFP permite estimar a produtividade do cafeeiro com até 6 meses de antecedência e com acurácia superior a 93% (FAHL; CARELLI, 2007). Desde então, alguns trabalhos têm utilizado o IFP para estimar a produtividade do cafeeiro (OLIVEIRA, 2007; MIRANDA; REINATO; SILVA, 2014; ROCHA et al., 2016).

O problema dos modelos existentes para estimativa da produtividade é a quantidade de variáveis necessárias e a dificuldade em mensurá-las. Os modelos agrometeorológicos e os agrometeorológicos-espectrais necessitam de informações de temperatura média do ar, velocidade do vento a 2 m de altura, umidade relativa do ar, insolação ou radiação solar, precipitação pluvial, altitude, capacidade de armazenamento de água disponível no solo e profundidade do sistema radicular (ROSA et al., 2010). Já o modelo com base no IFP necessita de uma mensuração da quantidade de frutos presentes em alguns nós produtivos dos ramos plagiotrópicos de várias plantas (FAHL et al., 2005). Essa tarefa pode ser lenta e fatigante, uma vez que o técnico responsável pelo levantamento fica exposto às condições de campo durante a coleta de dados.

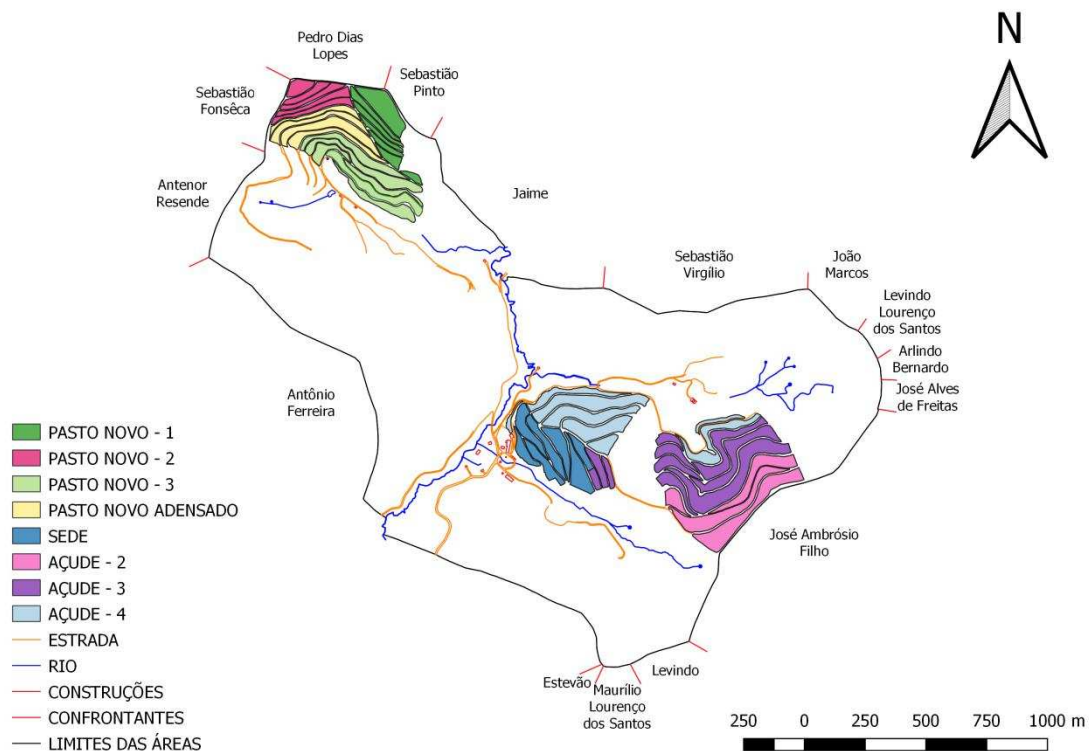
Diversos autores vêm aliando métodos de aprendizado de máquina (*machine learning*) na tentativa de estimar a produtividade de culturas agrícolas (PATEL; PATEL, 2016; GANDHI; ARMSTRONG, 2016; RAMESH; VARDHAN, 2015; MEDAR; RAJPUROHIT, 2014). *Machine learning* pode ser realizado com o auxílio de programas computacionais gratuitos. A resposta espectral da planta pode ser obtida por meio de imagens orbitais que, em alguns casos, também são gratuitas aos usuários. Logo, a existência de um modelo para estimativa de produtividade com base apenas na resposta espectral da cultura e em métodos de *machine learning* traria vantagem econômica aos usuários e aos produtores rurais. Dessa forma, objetivou-se desenvolver um sistema para estimativa de produtividade do café utilizando métodos de *machine learning* e índices de vegetação obtidos a partir de imagens orbitais.

3.2 MATERIAL E MÉTODOS

3.2.1 Área de estudo

O experimento foi conduzido em três áreas cultivadas com café da espécie *Coffea arábica* L., localizadas na Fazenda Braúna, município de Araponga, Minas Gerais (Figura 3.1). As três áreas cultivadas com café, juntas, totalizam 86 hectares e possuem 8 talhões. Esses talhões constituíram as unidades experimentais do trabalho. Como foram utilizadas a produtividades do café nos anos 2017 e 2018, as análises foram feitas no total de 16 unidades experimentais.

Figura 3.1. Área de estudo cultivada com café localizada na Fazenda Braúna município de Araponga, Minas Gerais, Brasil



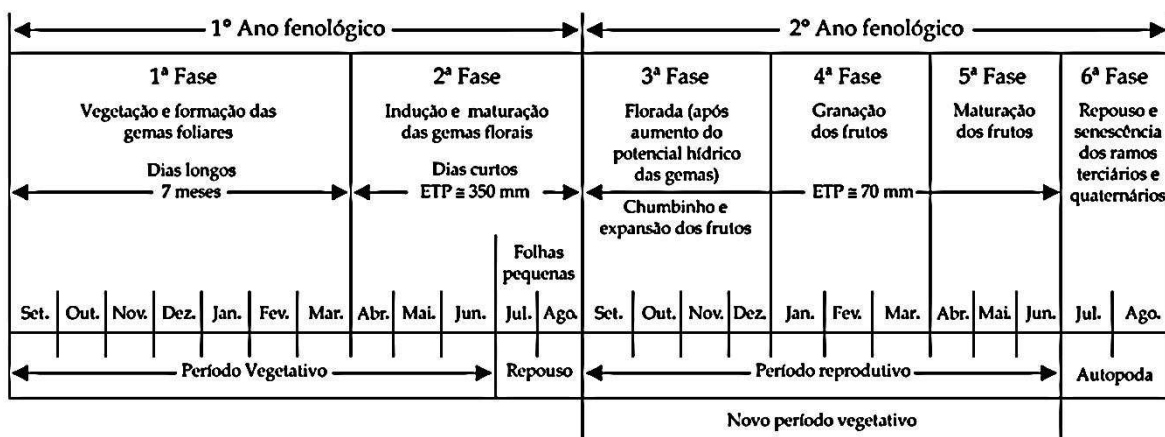
Fonte: Adaptado de Cerqueira (2004).

3.2.2 Imagens orbitais

Para compor o banco de dados com informações de todo o ciclo fenológico da cultura do café (Figura 3.2) foi necessário utilizar imagens oriundas de quatro satélites: *China-Brazil Earth Resources Satellite (Cbbers-4)*, *Land Remote Sensing Satellite (Landsat-8)*, *Resourcesat-2* e *Sentinel-2*. As imagens em que a área de

estudo esteve encoberta por nuvens foram descartadas. Cada imagem restante foi utilizada, após pré-processamento, para calcular os índices espectrais em cada uma das unidades experimentais presentes na área de estudo. As etapas que envolveram manipulação das imagens como recortes e reprojeções foram feitas no software QGIS versão 3.2.3 (QGIS Development Team, Open Source Geospatial Foundation, Chicago, IL, EUA).

Figura 3.2. Esquematização das seis fases fenológicas do cafeeiro arábica, durante 24 meses, nas condições climáticas tropicais do Brasil.



Fonte: CAMARGO; CAMARGO, 2001.

As imagens oriundas dos satélites Cbers-4 e Resourcesat-2 foram adquiridas no site da Divisão de Geração de Imagens do Instituto Nacional de Pesquisas Espaciais (INPE) enquanto as imagens do Landsat-8 foram baixadas do servidor da *United States Geological Survey* (USGS). Já as imagens do Sentinel-2 foram adquiridas no ESA portal online cujo nome é *Copernicus Open Access Hub*. As cenas de cada satélite englobaram a área contida na *path/row* correspondente a 217/074 do *Worldwide Reference System 2* (WRS-2) que é onde está localizada a sede da fazenda com as áreas de estudo. O sistema de coordenadas geográficas *World Geodetic System* (WGS), datum 84, foi adotado na realização deste trabalho.

3.2.3 Pré-processamento das imagens orbitais

A calibração radiométrica e a correção atmosférica corresponderam a primeira etapa de pré-processamento das imagens orbitais. A calibração radiométrica foi

realizada por meio da utilização das equações 3.1 e 3.2. Essas equações transformaram o nível digital das imagens em valores de radiância (Eq. 3.1) e em seguida em valores de reflectância (Eq. 3.2). Os parâmetros das equações foram obtidos nos metadados das imagens quando estas foram oriundas dos satélites Landsat-8 e Sentinel-2.

$$L_i(\lambda) = G(\lambda) \cdot DN(\lambda) + \text{offset}(\lambda) \quad \text{Eq. 3.1}$$

em que: $L_i(\lambda)$ é a radiância na banda λ no topo da atmosfera (top of atmosphere - TOA), $G(\lambda)$ é o coeficiente de ganho da banda λ ($\text{W} \cdot \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \mu\text{m}^{-1}$), $DN(\lambda)$ é o nível digital dos pixels da imagem e $\text{offset}(\lambda)$ é o coeficiente de viés para a banda λ ($\text{W} \cdot \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \mu\text{m}^{-1}$).

$$\rho_i(\lambda) = \frac{\pi \cdot L_i(\lambda) \cdot d^2}{E(\lambda)_{\text{SUN}} \cdot \cos(\theta_z)} \quad \text{Eq. 3.2}$$

em que: $\rho_i(\lambda)$ é a reflectância TOA na banda λ (adimensional), π é a constante matemática (adimensional), d é a distância Terra-Sol (unidades astronômicas), $E(\lambda)_{\text{SUN}}$ é a irradiância solar exoatmosférica ($\text{W} \cdot \text{m}^{-2} \cdot \text{m}^{-1}$) e θ_z é o ângulo zenital solar (radianos).

Na calibração das imagens do Cbers-4 e do Resourcesat-2, o coeficiente de ganho da banda λ foi calculado por meio da Eq. 3.3. Para isso, os dados de radiância máxima e mínima do sensor a bordo do Resourcesat-2 foram obtidos nos metadados da imagem. Já os dados de radiância máxima e mínima do sensor a bordo do Cbers-4, bem como o coeficiente de ganho da banda λ ($G(\lambda)$) e o coeficiente de viés para a banda λ ($\text{offset}(\lambda)$), foram obtidos nos trabalhos de Epiphanyo (2009) e Pinto et al. (2016). A irradiância solar exoatmosférica ($E(\lambda)_{\text{SUN}}$) nos sensores a bordo do Cbers-4 e do Resourcesat-2 foram obtidos nos trabalhos de Pinto et al (2016) e Keerthi e Kumar (2011), respectivamente. Após transformação dos níveis digitais em reflectância, foi utilizando o método *Dark Object Subtraction* (DOS) para realizar a correção atmosférica das imagens.

$$G(\lambda) = \frac{L(\lambda)_{\max} - L(\lambda)_{\min}}{Q(\lambda)_{cal\max} - Q(\lambda)_{cal\min}} \quad \text{Eq. 3.3}$$

em que: $L(\lambda)_{\max}$ é a radiância máxima na banda λ , $L(\lambda)_{\min}$ é a radiância mínima na banda λ , $Q(\lambda)_{cal\max}$ é a resolução radiométrica (quantização) máxima na banda λ do sensor, $Q(\lambda)_{cal\min}$ é a resolução radiométrica (quantização) mínima na banda λ do sensor.

Para que as imagens oriundas dos satélites Cbers-4, Landsat-8 e Resourcesat-2 pudessem ser utilizadas em conjunto foi realizada uma etapa de pré-processamento que consistiu em uma análise preditiva. As imagens originadas pelos sensores dos satélites foram utilizadas para prever imagens do satélite Sentinel-2. Isso significa que houve uma transformação das imagens Cbers-4, Landsat-8 e Resourcesat-2 em imagens Sentinel-2. Foram utilizados cinco modelos distintos para realizar essa transformação (Tabela 3.1).

Tabela 3.1. Modelos usados para transformar as imagens oriundas dos satélites Cbers-4, Landsat-8 e Resourcesat-2 em imagens Sentinel-2.

Banda predita do Sentinel-2	Satélite base da predição		
	Cbers-4	Landsat-8	Resourcesat-2
Azul	gaussprlinear	bridge	treebag
Verde	gaussprlinear	bridge	svmLinear
Vermelho	gaussprlinear	bridge	svmLinear
NIR	gaussprlinear	bridge	svmLinear
SW1	svmLinear	bridge	svmLinear
SW2	treebag	bridge	svmLinear

svmLinear é o método de *machine learning* máquinas de vetores de suporte com Kernel linear; treebag é a árvore de classificação e regressão com agregação Bootstrap bridge corresponde à regressão ridge Bayesiana; e gaussprLinear é o método de *machine learning* processo Gaussiano. NIR é a banda espectral do infravermelho próximo, SW1 e SW2 são as bandas espectrais do infravermelho de ondas curtas um e dois, respectivamente.

A correção topográfica das imagens foi realizada por meio do uso do método empírico-rotacional (TAN et al., 2010; TAN et al., 2013). A primeira etapa desse método consiste no cálculo do ângulo de incidência da radiação na superfície. A Eq. 3.4 foi utilizada com esse propósito.

$$IL = \cos Z \cdot \cos S + \text{sen}Z \cdot \text{sen}S \cdot \cos(\varphi_Z - \varphi_S) \quad \text{Eq. 3.4}$$

em que: IL é o ângulo de incidência da radiação na superfície, Z é o ângulo zenital solar, S é a declividade do terreno, φ_Z é o ângulo azimutal solar e φ_S é o aspecto do terreno. IL varia de -1 (iluminação mínima) a 1 (iluminação máxima).

A etapa seguinte do procedimento de correção topográfica foi a modelagem utilizando análise de regressão linear entre a reflectância na superfície inclinada e o ângulo de incidência solar (IL). Essa regressão linear (Eq. 3.5) forneceu os valores do coeficiente angular da reta (parâmetro a) e o intercepto da reta no eixo Y (parâmetro b). Como última etapa do procedimento, a Eq. 3.6 foi utilizada para realizar a correção topográfica das imagens.

$$\rho_i(\lambda) = a \cdot IL + b \quad \text{Eq. 3.5}$$

em que: $\rho_i(\lambda)$ é a reflectância em uma superfície inclinada, IL é o ângulo de incidência solar na superfície e a e b são os parâmetros do modelo linear ajustado.

$$\rho_h(\lambda) = \rho_i(\lambda) - (a \cdot IL + b) \quad \text{Eq. 3.6}$$

em que: $\rho_h(\lambda)$ é a reflectância corrigida, $\rho_i(\lambda)$ é a reflectância em uma superfície inclinada, IL é o ângulo de incidência solar na superfície e a e b são os parâmetros da equação da reta.

3.2.4 Índices espectrais

Após o processamento digital das imagens orbitais, foram determinados os

índices: *Normalized Difference Vegetation Index* (NDVI); *Normalized Difference Water Index* (NDWI); *Soil Adjusted Vegetation Index* (SAVI); *Green Vegetation Index* (GVI); *Weighted Difference Vegetation Index* (WDVI); e o *tasseled cap wetness index* (WETNESS). As reflectâncias nas bandas do azul, verde, vermelho, infravermelho próximo, infravermelho de ondas curtas 1 (*Shortwave Infrared 1 – SWIR 1*) e infravermelho de ondas curtas 2 (SWIR 2) consideradas no cálculo dos índices (Tabela 3.2) foram representadas como B2, B3, B4, B5, B6, e B7 , respectivamente.

Tabela 3.2. Métodos para calcular os índices espectrais.

Índices	Método de cálculo	Referência
NDVI	$NDVI = \frac{(B5 - B4)}{(B5 + B4)}$	SONG et al. (2016)
SAVI	$SAVI = (1 - L) \times \frac{(B5 - B4)}{(B5 + B4 + L)}$	HUETE (1988)
WDVI	$WDVI = B5 - Sls \times B4$	RICHARDSON e WIEGAND(1977)
NDWI	$NDWI = \frac{(B5 - B6)}{(B5 + B6)}$	GAO (1996)
WETNESS	$WETNESS = 0,2626 \cdot B2$ $+ 0,2141 \cdot B3$ $+ 0,0926 \cdot B4$ $+ 0,6560 \cdot B5$ $- 0,7629 \cdot B6$ $- 0,5388 \cdot B7$	HUANG et al. (2002)
GVI	$GVI = -0,2848 \cdot B2$ $- 0,2435 \cdot B3$ $+ 0,5436 \cdot B4$ $+ 0,7243 \cdot B5$ $+ 0,0840 \cdot B6$ $- 0,1800 \cdot B7$	MATHER (1999)

* "B" refere-se às reflectâncias nas bandas das imagens orbitais; "L" = 0,5; Linha de solo e declive foram definidos com base na relação da reflectância do solo entre as bandas B3 e B4; Sls é a abreviatura da expressão *Soil line slope*. Fonte: Adaptado de SATIR e BERBEROGLU (2016).

Os índices foram calculados para diferentes fases de desenvolvimento da cultura do café (Figura 3.2). A época de análise dos índices compreendeu o intervalo de tempo correspondente aos ciclos fenológicos das colheitas dos anos de 2017 e 2018. Ou seja, foram utilizadas imagens de setembro de 2015 até a colheita da safra 2016/2017 (maio até agosto de 2017) e até a colheita da safra 2017/2018 (maio até agosto de 2018).

Após o cálculo dos índices NDVI, SAVI, WDVI, NDWI, WETNESS, GVI, foram traçados oito perfis temporais de cada um deles e para todas as unidades experimentais. O primeiro, segundo, terceiro, quarto, quinto e sexto perfil temporal corresponderam, respectivamente, ao período da 1^a, 2^a, 3^a, 4^a, 5^a, e 6^a fase fenológica do cafeeiro (Figura 3.2). O sétimo e o oitavo perfil temporal foram referentes, respectivamente, ao ano 1 e ao ano 2 do ciclo fenológico da cultura.

Em cada fase do ciclo fenológico, foi ajustado um modelo por meio de regressão linear simples entre os valores de cada índice em cada unidade experimental. O valor do coeficiente angular do modelo linear ajustado foi inserido no banco de dados junto ao valor médio de cada índice. Esse procedimento também foi realizado nas bandas das imagens. Ou seja, o banco de dados ficou composto pelos valores do coeficiente angular e da média dos valores dos índices e também das bandas das imagens.

3.2.5 Modelo digital de elevação

O modelo digital de elevação (MDE) foi gerado pelo programa computacional QGIS versão 3.2.3 (QGIS Development Team, Open Source Geospatial Foundation, Chicago, IL, EUA). Para isso, foram utilizados dados de levantamento topográfico, obtidos por meio de um aparelho DGPS da marca Trimble, modelo Pro XT. A correção diferencial foi realizada utilizando o Pathfinder Office® v. 5.00, fornecido pelo fabricante do aparelho DGPS. Para a correção diferencial foram utilizados dados da base do Instituto Brasileiro de Geografia e Estatística (IBGE), localizada no município de Viçosa - MG. A partir da análise do modelo digital de elevação, no programa computacional QGIS, foram obtidas informações sobre a altitude (m) mínima, média e máxima, declividade (°) mínima, média e máxima e aspecto (°) mínimo, médio e máximo do terreno em cada unidade experimental. Essas

informações serviram como variáveis de entrada no modelo de estimativa de produtividade.

3.2.6 Produtividade da cultura

A produtividade da cultura foi estimada por meio do índice fenológico de produção (IFP) como proposto por Fahl et al. (2005). Para isso, foram amostradas 60 plantas de café em cada unidade experimental. A amostragem foi realizada em diferentes entrelinhas da cultura, escolhidas aleatoriamente. Em cada planta foi contado o número de frutos presentes no 4° e 5° nós produtivos de 2 ramos plagiotrópicos, um em cada lado da planta. Foram medidas, ainda, em cada entrelinha, as alturas de cinco plantas para obtenção da altura média das plantas de cada unidade experimental. Esses dados foram utilizados para cálculo do IFP por meio da Eq. 3.7 (ALFONSI, 2008) que por sua vez foi utilizado para estimar a produtividade por meio da Eq. 3.8 (FAHL et al., 2005; OLIVEIRA, 2007).

$$IFP = \frac{10000}{ESP} \cdot 2 \cdot ALT \cdot NF_{45} \quad \text{Eq. 3.7}$$

em que: *IFP* é o índice fenológico de produção; *ESP* é o espaçamento entre linhas de cultivo; *ALT* é a altura média das plantas; *NF₄₅* é a média dos números de frutos dos 4° e 5° nós produtivos.

$$PROD = 0,0005 \cdot IFP \quad \text{Eq. 3.8}$$

em que: *PROD* é a produtividade estimada.

A produtividade da cultura foi determinada, ainda, por meio do método tradicional que consistiu na contagem da quantidade de balaios do material colhido para uma determinada área nos anos de 2017 e 2018. A área em questão foi cada unidade experimental (talhões da fazenda). Isso proporcionou a comparação entre os resultados dos modelos propostos e os resultados da metodologia de estimativa da produtividade com base no IFP.

3.2.7 Análise dos dados e modelo para estimativa de produtividade

A transformação dos dados em conhecimento aconteceu por meio do uso de métodos de aprendizado de máquina (*machine learning*). Para isso, as unidades experimentais (talhões da fazenda Braúna) foram selecionadas e separadas nos conjuntos de dados para treinamento, teste e avaliação do sistema para estimativa de produtividade do café. Cada unidade experimental foi caracterizada por variáveis de entrada e saída. Todas as variáveis foram normalizadas para diminuir o efeito da escala na modelagem. A variável de saída dos modelos foi a produtividade das unidades experimentais em $\text{litros.hectare}^{-1}$. As variáveis de entrada (também denominadas de variáveis preditoras) dos modelos foram:

- Os valores do coeficiente angular de cada índice para cada perfil temporal (adimensional);
- Os valores do coeficiente angular de cada banda para cada perfil temporal (adimensional);
- Os valores médios de cada índice para cada perfil temporal (adimensional);
- Os valores médios de reflectância em cada banda espectral para cada perfil temporal (adimensional);
- Densidade de plantas na unidade experimental ($\text{número de plantas.hectare}^{-1}$);
- Altitude mínima na unidade experimental (m);
- Altitude média na unidade experimental (m);
- Altitude máxima na unidade experimental (m);
- Declividade do terreno mínima na unidade experimental ($^{\circ}$);
- Declividade do terreno média na unidade experimental ($^{\circ}$);
- Declividade do terreno máxima na unidade experimental ($^{\circ}$);
- Aspecto do terreno mínimo na unidade experimental ($^{\circ}$);
- Aspecto do terreno médio na unidade experimental ($^{\circ}$);
- Aspecto do terreno máximo na unidade experimental ($^{\circ}$);

3.2.8 Correlação entre as variáveis preditoras

A correlação de Pearson foi utilizada para analisar a correlação entre as variáveis preditoras dos modelos (variáveis de entrada). Uma das variáveis que apresentou correlação maior do que 0,7 foi excluída do conjunto de treinamento e teste dos modelos. Dessa forma, das 202 possíveis variáveis preditoras, restaram apenas 14 variáveis cuja correlação entre si foi menor do que 0,7.

3.2.9 Treinamento dos modelos de estimativa da produtividade

Os métodos de regressão linear (lm), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel*- svmLinear1), redes neurais artificiais (nnet1), regressão ridge Bayesiana (*Bayesian Ridge Regression* - bridge), processo Gaussiano (*Gaussian Process* - gaussprLinear) e floresta aleatória (*random forest*) foram usados para estimar a produtividade do cafeeiro. O valor um foi usado no parâmetro custo método svmLinear1. As redes neurais artificiais foram treinadas com um neurônio na camada intermediária da rede. O valor um foi usado no parâmetro decaimento do peso das redes neurais artificiais. Dessa forma, seis métodos de *machine learning* foram utilizados para estimar a produtividade do cafeeiro. Os pacotes e os parâmetros de treinamento de cada método são apresentados na Tabela 3.3.

Tabela 3.3. Pacotes na linguagem R utilizados para executar os métodos de *machine learning* e parâmetros de treinamento de cada método.

Nome	Representação do método	Biblioteca	Parâmetros de ajuste
Regressão linear	lm	nativa do R	
máquinas de vetores de suporte com Kernel linear (<i>support vector machine with linear Kernel</i>)	svmLinear1	e1071	custo = 1
redes neurais artificiais	nnet1	nnet1	número de neurônios na camada intermediária da rede = 1; decaimento do peso = 1
<i>Bayesian Ridge Regression</i>	bridge	monomvn	
<i>Gaussian Process</i>	gaussprLinear	caret	
<i>Random forest</i>	randomForest	randomForest	

Os métodos foram executados no programa computacional R Versão 3.5.1 (R Team, 2018). O pacote Caret (*Classification And Regression Training*) foi utilizado como interface para simplificar o processo de criação dos modelos e visualização dos resultados. Duas funções foram usadas para controlar a aleatoriedade envolvida na modelagem e garantir resultados reproduzíveis. A primeira função foi a *set.seed*. A outra função foi a *trainControl* que criou um objeto de controle. Todos os modelos realizaram a modelagem com base nesse mesmo objeto, ou seja, a modelagem pelos distintos métodos de mineração foi feita com os mesmos dados de treinamento e teste. Esse fato permitiu a comparação entre os resultados apresentados pelos métodos.

O objeto de controle foi criado a partir da utilização do método de validação cruzada *leave one out*. Esse método consistiu em treinar o modelo deixando um dado de fora do treinamento a cada iteração do algoritmo. O dado que ficou de fora do treinamento foi usado para testar a estimativa. De maneira iterativa, cada subconjunto foi utilizado uma vez como conjunto de teste. O algoritmo calculou o

erro médio absoluto (*mean absolute error* - MAE) entre o valor real e o valor estimado pelo modelo. O cálculo do erro foi realizado utilizando o dado reservado para teste.

O treinamento foi realizado a partir da utilização de distintos conjuntos de variáveis de entrada. O primeiro conjunto para treinamento utilizou 14 variáveis de entrada. O segundo conjunto de treinamento utilizou apenas as 10 principais variáveis resultantes da função *varImp*. Essa função implementou um método genérico para calcular a importância de variáveis em modelos de regressão e classificação para modelos produzidos pela função *train* (Kuhn et al., 2019).

O terceiro e quarto conjuntos de variáveis corresponderam as 10 principais variáveis obtidas por meio da análise com o algoritmo *randomForest*. O algoritmo *randomForest* determinou a importância das variáveis com base em duas métricas características de modelagens do tipo regressão. A primeira métrica correspondeu ao incremento no erro quadrado médio (*mean squared error* – MSE) quando a modelagem foi realizada sem uma das variáveis de entrada (Breiman et al., 2018). A segunda métrica foi a diminuição total nas impurezas do nó (Node Purity) nas divisões das árvores de decisão que compuseram o modelo (Breiman et al., 2018). A pureza dos nós foi medida pela soma residual dos quadrados, sendo esta uma característica das modelagens do tipo regressão (Breiman et al., 2018). O quinto, sexto e sétimo conjuntos de variáveis utilizaram as 5 variáveis mais importantes filtradas por meio dos métodos genérico, MSE e Node Purity.

3.2.10 Avaliação dos modelos de estimativa da produtividade

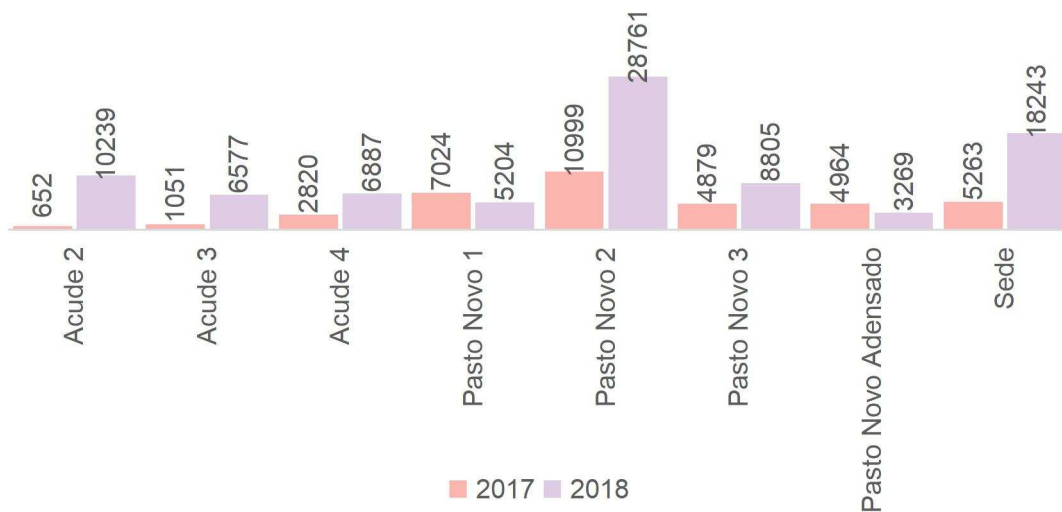
Os valores da produtividade estimados pelo algoritmo, bem como os valores estimados pelo método do índice fenológico de produção para cada unidade experimental, foram comparados com os valores da produtividade da cultura determinada por meio do método tradicional. O erro absoluto médio (MAE) foi utilizado para avaliar a acurácia dos modelos e a magnitude dos erros na estimativa da produtividade. Os dados usados na etapa de avaliação dos modelos foram referentes aos talhões Açude 4 e Sede no ano de 2018.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Produtividade de café

As produtividades nos talhões colhidas nos anos de 2017 e 2018 são apresentadas na Figura 3.3. A maior produtividade ocorreu no talhão Pasto Novo2 (PN2) em 2018. Com exceção do talhão Pasto Novo 1 (PN1) e Pasto Novo Adensado (PNA), todos os demais talhões apresentaram uma produtividade maior na colheita de 2018 do que em 2017. Nesses talhões, a produtividade média em 2018 foi 310% maior do que a produtividade média em 2017. As variações nas produtividades de um ano para o outro, além de outros fatores, podem ter sido causadas pela bienalidade do café, conforme relatado por Rosa et al. (2010) e Silva e Reis (2013).

Figura 3.3. Produtividade real nas unidades experimentais em litros.hectare⁻¹.



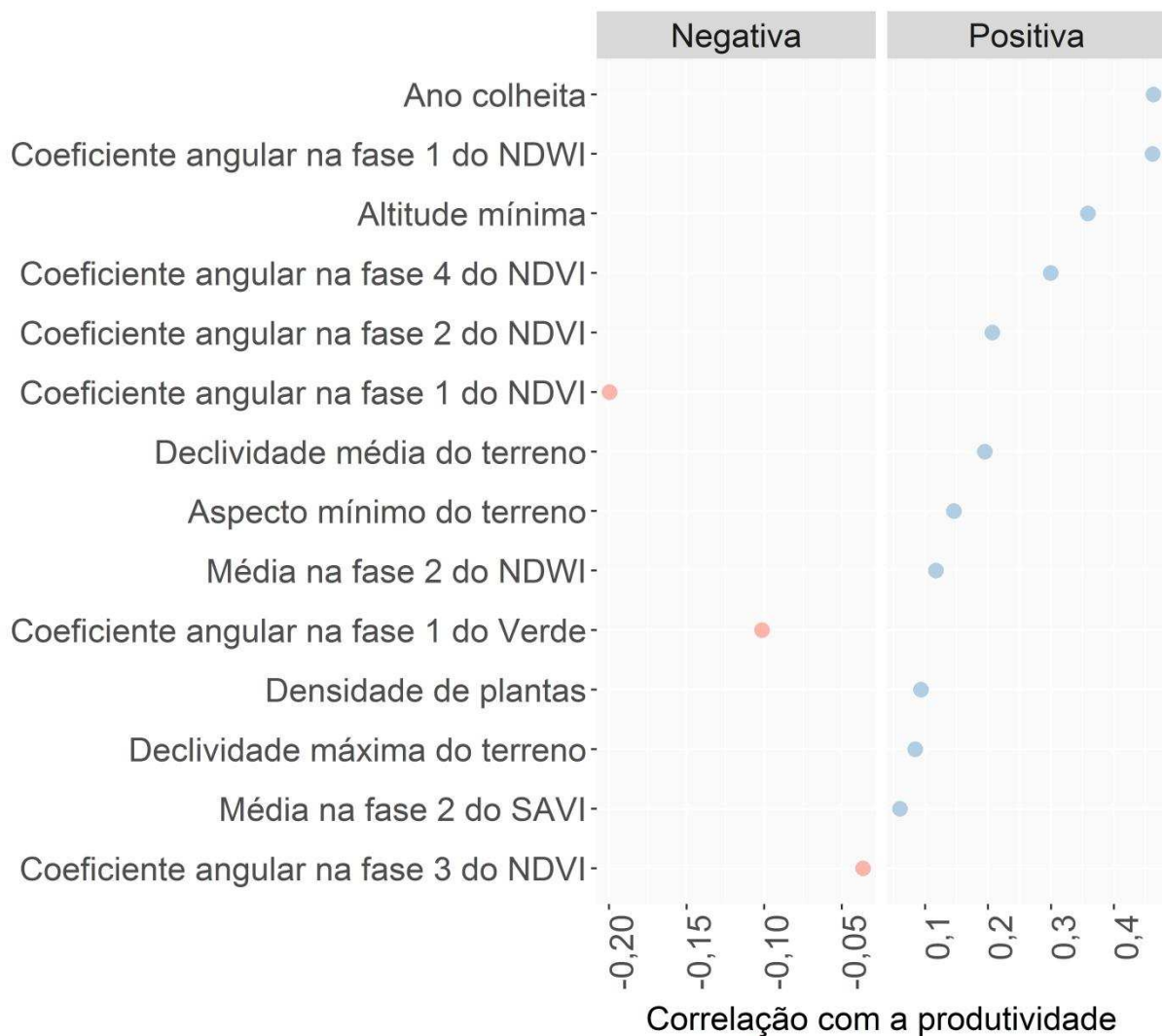
3.3.2 Correlação entre as variáveis preditoras

Na Figura 3.4 são apresentadas as variáveis que apresentaram correlação entre si menor do que 0,7. Após a exclusão das variáveis que apresentaram correlação maior do que 0,7 umas com as outras, restaram apenas 13 variáveis das 202 variáveis iniciais. Além do ano da colheita, essas 13 variáveis foram as variáveis preditoras utilizadas nos modelos de predição da produtividade do cafeeiro.

3.3.3 Importância das variáveis de entrada nos modelos de estimativa da produtividade de café

Na Figura 3.5 são apresentadas as correlações entre a produtividade do cafeeiro e as variáveis de entrada dos modelos de estimativa da produtividade. O ano da colheita foi a variáveis que apresentou maior correlação com a produtividade do cafeeiro. A variável que apresentou a segunda maior correlação com a produtividade foi o coeficiente angular na fase 1 do NDWI.

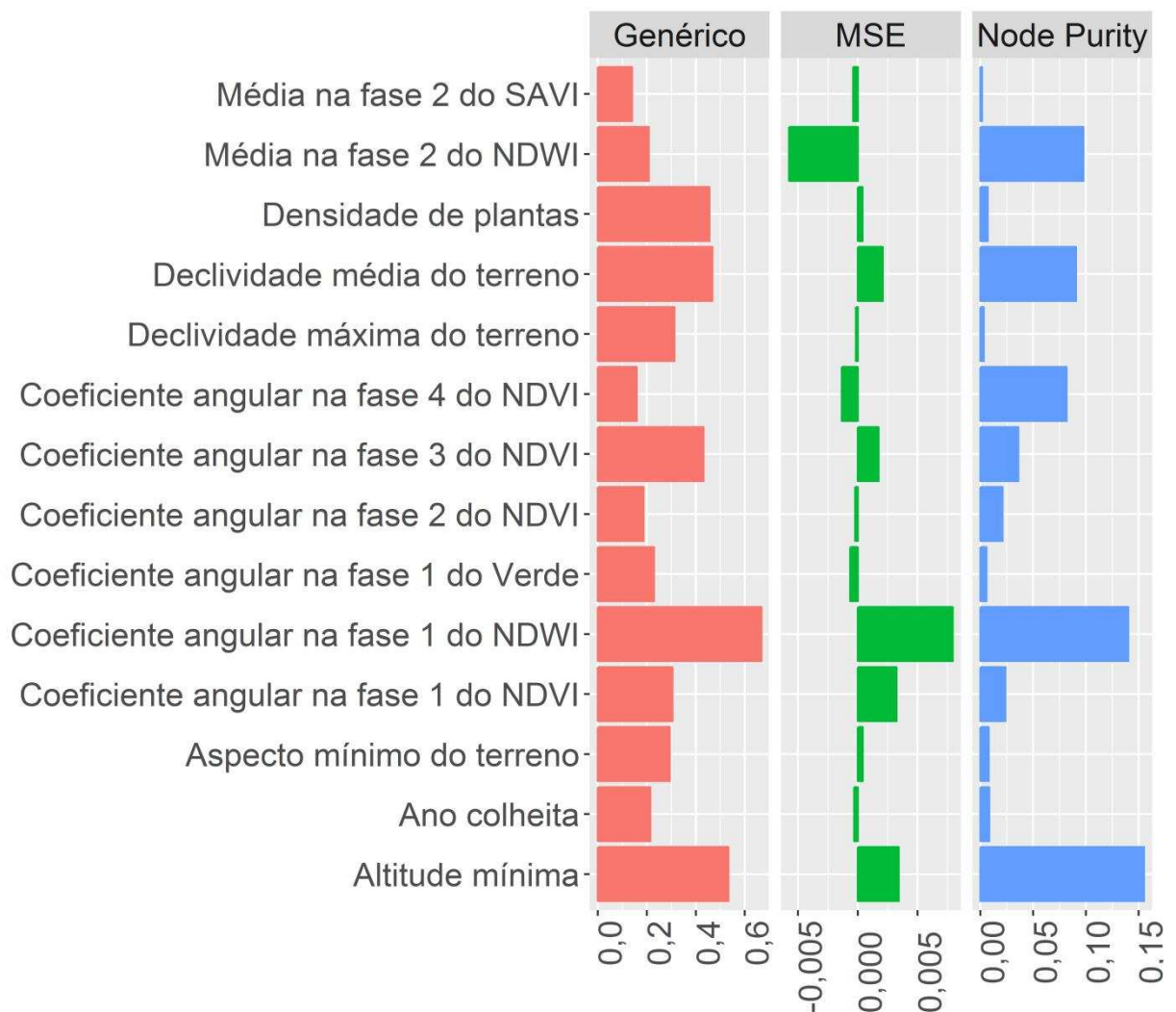
Figura 3.5. Correlações entre a produtividade do cafeeiro e as variáveis de entrada dos modelos de estimativa da produtividade



O coeficiente angular na fase 1 do NDWI foi a variável mais importante na

estimativa da produtividade do cafeeiro pelo método genérico e pelo método do MSE (Figura 3.6). essa variável foi a segunda mais importante pelo método Node Purity. A altitude mínima do terreno foi a variável mais importante pelo método Node Purity e a segunda variável mais importante pelo método genérico e pelo método do MSE. Nota-se, então, que essas duas variáveis foram as que mais contribuíram na estimativa da produtividade agrícola do cafeeiro conforme indicou os três métodos de determinação da importância das variáveis.

Figura 3.6. Importância das variáveis determinada pelos métodos: Genérico; randomForest: MSE; e randomForest : Node Purity.



O coeficiente angular na fase 1 do NDWI e a altitude mínima do terreno apresentaram, respectivamente, a segunda e a terceira maior correlação com a

produtividade do café (Figura 3.5). O ano da colheita foi a variável que apresentou a maior correlação com a produtividade do cafeeiro. No entanto, essa variável não foi considerada como uma das mais importantes pelo método genérico e nem pelos métodos do MSE e Node Purity. Ou seja, a correlação com a variável estimada pelos modelos de estimativa não necessariamente indica que a variável preditora é importante ou não na estimativa.

3.3.4 Teste dos modelos de estimativa da produtividade de café

A raiz do erro quadrático médio (*root mean square error* - RMSE) resultante dos modelos treinados é apresentada na Tabela 3.1. O RMSE mínimo ocorreu na modelagem que utilizou o método de máquinas de vetores de suporte com Kernel linear (*support vector machine with linear Kernel* - svmLinear1) com 5 variáveis filtradas por meio do método genérico de determinação da importância das variáveis. O valor do RMSE mínimo foi de 6530 litros de café.hectare⁻¹ que correspondeu a 118,7% de diferença média entre os valores estimados e a produtividade real.

Tabela 3.1. Raiz do erro quadrático médio (*root mean square error* - RMSE) em litros.hectare⁻¹ para todos os modelos treinados para estimar a produtividade

Métodos	14 variáveis	5 variáveis			10 variáveis		
		Genérico	MSE	Node Purity	Genérico	MSE	Node Purity
lm	79189 ^a	7801 ^b	7417 ^b	7904 ^b	10945 ^b	11668 ^b	10125 ^b
svmLinear1		6530 ^b	6611 ^b	7411 ^b			
nnet1	8404 ^b	8405 ^b	8404 ^b	8406 ^b	8404 ^b	8405 ^b	8405 ^b
bridge		7869 ^b	7791 ^b	7837 ^b			
gaussprLinear		7087 ^b	6934 ^b	7506 ^b			
randomForest	8388 ^b	7237 ^b	7026 ^b	8521 ^b	7474 ^b	7817 ^b	8400 ^b

O valor do RMSE é referente à média dos erros na estimativa da produtividade nos 16 talhões da propriedade. Isto é, para chegar nesse valor médio, as estimativas em alguns talhões apresentaram erros maiores e alguns talhões

apresentaram erros menores. Dentre as estimativas nos talhões, a que apresentou o menor erro foi a estimativa da produtividade do talhão Pasto Novo 1 no ano de 2017. Nesse caso, o erro correspondeu a uma diferença de 11% entre o valor estimado e o valor observado da produtividade.

Os erros obtidos nas modelagens realizadas no presente estudo foram superiores aos erros relatados por Rosa et al. (2010) que variaram de 0,4% a 8,5% e Aparecido et al. (2017) que variaram de 0,57% a 0,90%. A diferença entre os erros obtidos no atual trabalho e os erros obtidos nos estudos realizados por esses pesquisadores pode ter sido causada pelo tipo de informação utilizada para criar os modelos de estimativa da produtividade. Enquanto Rosa et al (2010) e Aparecido et al. (2017) utilizaram dados com informações agrometeorológicas e espectrais em suas modelagens, no presente estudo os dados utilizados na modelagem eram compostos, em maior proporção, apenas por informações espectrais da planta. Com isso, tanto o tipo quanto o número de informações utilizadas podem ter influenciado na acurácia dos modelos obtidos.

As diferenças de RMSE's apresentados pelos modelos e ilustrados na Tabela 3.1 foram suficientes para classificar os modelos em dois grupos pelo teste do Scott-Knot. Apenas o modelo que utilizou o método de regressão linear com 14 variáveis preditoras foi diferente dos demais. Os modelos classificados no grupo b são todos semelhantes entre si.

O erro médio absoluto (*mean absolute error* - MAE) resultante dos modelos treinados é apresentado na Tabela 3.2. O MAE mínimo foi de 4623 litros de café.hectare⁻¹. Esse erro correspondeu a 74,7% de diferença média entre os valores estimados e a produtividade real. Diferente do RMSE mínimo, o MAE mínimo ocorreu na modelagem que utilizou o método randomForest com 5 variáveis cuja filtragem ocorreu pela técnica MSE do algoritmo randomForest. Com exceção do modelo que utilizou a regressão linear com 14 variáveis para estimar a produtividade, todos os demais modelos apresentaram resultados semelhantes entre si pelo teste do Scott-Knott.

Tabela 3.2. Erro médio absoluto (*mean absolute error* - MAE) para todos os modelos treinados para estimar a produtividade

Métodos	14 variáveis	5 variáveis			10 variáveis		
		Genérico	MSE	Node Purity	Genérico	MSE	Node Purity
lm	36669 ^a	5911 ^b	5719 ^b	6237 ^b	8529 ^b	8839 ^b	7744 ^b
svmLinear1		4926 ^b	4774 ^b	5670 ^b			
nnet1	7153 ^b	7158 ^b	7157 ^b	7161 ^b	7155 ^b	7155 ^b	7157 ^b
bridge		5670 ^b	5672 ^b	5668 ^b			
gaussprLinear		5381 ^b	5118 ^b	5820 ^b			
randomForest	6143 ^b	5441 ^b	4623 ^b	5904 ^b	5263 ^b	5713 ^b	6084 ^b

Da mesma forma que o RMSE, o valor do MAE é referente à média dos erros na estimativa da produtividade nos 16 talhões da propriedade. Dentre as estimativas nos talhões, a que apresentou o menor erro foi a estimativa da produtividade do talhão Açude 3 no ano de 2018. Nesse caso, o erro correspondeu a uma diferença de 1,7% entre o valor estimado e o valor observado da produtividade.

Petersen (2018) conseguiu estimar a produtividade da cultura do milho, da soja e do sorgo com erros médios de 5,7%, 5,8% e 22%, respectivamente. A autora considerou que seu modelo correspondeu a um método simples e bom indicador da produtividade das culturas. Considerando o erro médio, esses erros foram menores do que os erros apresentados no atual trabalho. É possível que as diferenças entre os erros encontrados no atual trabalho e no trabalho de Petersen (2018) sejam devido ao comportamento das culturas perenes (café no atual trabalho) e das culturas anuais (milho, soja e sorgo no trabalho de Petersen (2018)). A metodologia utilizada nos dois trabalhos foi diferente e isso também pode ter levado às diferenças nos resultados dos dois trabalhos.

Outra possibilidade do erro do atual trabalho ter sido maior é devido à bienalidade do café. Porque os perfis espectrais podem ser semelhantes nos anos de baixa e alta produtividade. Dessa forma, o modelo fundamentado apenas em informações espectrais pode confundir se aquela informação espectral correspondeu a um ano de maior ou menor produtividade. Para contornar esse problema, Aparecido et al. (2017) criaram um modelo para os anos de alta produtividade e um modelo para os anos de baixa produtividade. O fato da lavoura de café estudada

neste trabalho ter sido previamente recepada em alguns talhões e esqueletada em outros também pode ter influenciado os resultados encontrados no atual trabalho, pois a resposta espectral da cultura podia estar mais relacionada ao crescimento vegetativo devido à sua recuperação do que com o crescimento reprodutivo.

A estimativa da produtividade por meio da utilização de cinco variáveis foi semelhante às estimativas com os distintos métodos utilizando 10 ou 14 variáveis de entrada nos modelos. Dessa forma, a modelagem com cinco variáveis se destacou por corresponder a modelos mais simples que ofereceram resultados parecidos com os resultados dos modelos mais complexos. Nas cinco variáveis mais importantes, estão presentes características do terreno como a altitude mínima e a declividade média além de informações espectrais como o NDWI e o NDVI, principalmente na fase 1 e fase 3 do ciclo produtivo do café. Ou seja, como a fase 1 e a fase 3 do ciclo ocorrem, respectivamente, 12 e 3 meses antes do início da colheita seria possível realizar a estimativa da produtividade com 3 meses de antecedência.

Shiu e Chuang (2019) estimaram a produtividade de arroz na região central de Taiwan com erro percentual entre 0,06% a 13,22%. As variáveis preditoras utilizadas por estes autores incluíram quatro bandas originais, 11 índices de vegetação e 32 índices de textura derivadas de imagens orbitais. As variáveis que promoveram o menor erro foram a banda do azul, o índice de verde (*Greenness index*), o SAVI modificado e o SAVI, a entropia na banda NIR e a média do NIR. Semelhante a este trabalho, o estudo de Shiu e Chuang (2019) mostrou que o valor original da banda espectral pode ser útil na estimativa da produtividade, pois o coeficiente angular na fase 1 da banda do verde foi uma das 10 variáveis mais importantes pelo método genérico. O SAVI não foi uma das variáveis que mais contribuiu na estimativa da produtividade do cafeeiro, tanto no trabalho atual quanto no trabalho de Shiu e Chuang (2019). No entanto, no atual trabalho, isso ocorreu devido a sua alta correlação com as demais variáveis de entrada dos modelos, o que fez com que sua presença na modelagem não fosse necessária, pois outras variáveis com alta correlação já estavam sendo consideradas. A diferença entre os resultados dos dois trabalhos foi o fato de que Shiu e Chuang (2019) utilizaram variáveis relacionadas à textura enquanto neste trabalho elas não foram incluídas.

Na Figura 3.7 é apresentada a produtividade real observada nos talhões e a produtividade estimada pelos métodos de *machine learning*. Os métodos de *machine learning* bridge e redes neurais artificiais estimaram quase o mesmo valor

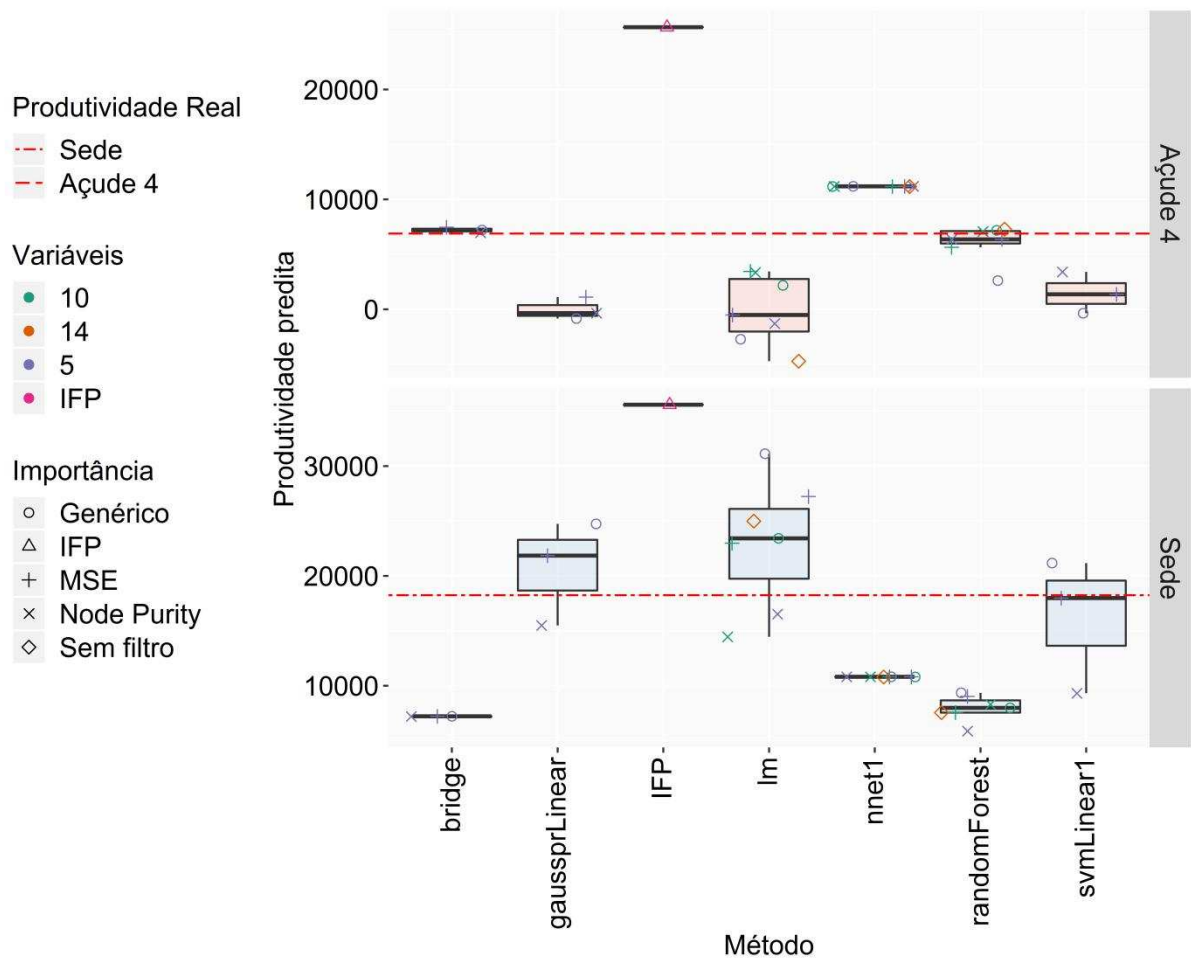
para as distintas produtividades observadas. Na Figura 3.7, essa situação é ilustrada por uma reta vertical. Os métodos `gaussprLinear`, `svmLinear1` e `bridge` só conseguiram criar o modelo quando foram utilizadas 5 variáveis preditoras. Nos métodos regressão linear e `randomForest`, houve uma tendência do ajuste com 5 variáveis ter apresentado menor dispersão entre os valores estimados e os valores observados da produtividade do cafeeiro.

nas impurezas do nó (Node Purity) do algoritmo randomForest. Os métodos de *machine learning* são: regressão ridge Bayesiana (bridge), processo Gaussiano (gaussprLinear), regressão linear (lm), redes neurais artificiais com um neurônio na camada intermediária (nnet1), floresta aleatória (randomForest) e máquinas de vetores de suporte com Kernel linear com função de custo igual a um (svmLinear1).

3.3.5 Avaliação dos modelos de estimativa da produtividade de café

A produtividade de café estimada pelos modelos é apresentada na Figura 3.8. A produtividade estimada pelo método do índice fenológico de produção (IFP) foi maior do que a produtividade observada nos talhões Açude 4 e Sede. Essa estimativa foi a maior dentre todos os métodos de estimativa da produtividade analisados neste trabalho.

Figura 3.8. Produtividades estimadas por diversos modelos.



As diferenças entre os valores estimados e a produtividade real nos talhões Açude 4 e Sede geraram os valores da raiz do erro quadrático médio (*root mean*

square error - RMSE) e do erro médio absoluto (*mean absolute error* - MAE) apresentados na Tabela 3.3. O método embasado no IFP estimou a produtividade do cafeeiro com um RMSE de 18069 litros.hectare⁻¹ e um MAE de 18055 litros.hectare⁻¹. Todos os erros obtidos pelos modelos utilizados neste trabalho foram menores do que os erros obtidos pelo método do IFP.

Tabela 3.3. Raiz do erro quadrático médio (*root mean square error* - RMSE) e erro médio absoluto (*mean absolute error* - MAE) das estimativas

Métodos	14 variáveis	5 variáveis			10 variáveis			IFP
		Genérico	MSE	Node Purity	Genérico	MSE	Node Purity	
RMSE								
lm	10147	12016	8889	6576	5604	4790	4323	
svmLinear1		6183	4572	7440				
nnet1	6730	6719	6718	6720	6725	6724	6725	
bridge		8455	8467	8461				
gaussprLinear		7798	5470	6133				
randomForest02	8217	7638	7178	9415	7914	8264	7683	
IFP								18069
MAE								
lm	9828	11900	8852	5615	5599	4741	4321	
svmLinear1		5739	3554	6872				
nnet1	6519	6512	6511	6514	6516	6516	6517	
bridge		6325	6453	6213				
gaussprLinear		7770	5345	5653				
randomForest	6199	7247	5524	7113	5928	6626	5713	
IFP								18055

A escolha das plantas para coleta dos frutos e posterior cálculo do IFP seguiu o esquema de amostragem aleatória. No momento da contagem dos frutos nas plantas selecionadas, foi observado que havia uma concentração dos frutos em algumas regiões da planta. Esse fato pode ter sido resultado das práticas culturais realizadas anteriormente na cultura como o esqueletamento e recepa. Como os frutos não estavam distribuídos uniformemente em toda a planta, a determinação da produtividade pelo IFP pode ter sido prejudicada. Afinal, a concentração de frutos nos locais aos quais os frutos deveriam ser contabilizados não correspondia à mesma distribuição de frutos na planta inteira.

O RMSE e o MAE mínimo foram resultantes da estimativa por meio dos métodos `lm` e `svmLinear1`, respectivamente. No caso do método `lm`, o ajuste com o RMSE mínimo ocorreu quando foram utilizadas 10 variáveis filtradas pelo método Node Purity. Quanto ao `svmLinear1`, o ajuste com MAE mínimo ocorreu quando foram utilizadas as cinco principais variáveis indicadas pelo método do incremento no MSE.

3.4 CONCLUSÃO

Os métodos de *machine learning* máquinas de vetores de suporte com Kernel linear com função de custo igual a um (`svmLinear1`), redes neurais artificiais com um neurônio na camada intermediária (`nnet1`), regressão ridge Bayesiana (`bridge`) e processo Gaussiano (`gaussprLinear`) apresentaram erros RMSE e MAE da estimativa de produtividade semelhantes uns aos outros pelo teste de Scott-Knott.

Foi possível estimar a produtividade a partir da utilização de cinco variáveis com erros semelhantes, a 5% de significância pelo teste de Scott-Knott, aos erros apresentados pelos modelos com 10 e com 14 variáveis que incluíam informações espectrais, topográficas e agronômicas.

A estimativa da produtividade pode ser realizada com até três meses de antecedência.

Todos os modelos compostos por métodos de *machine learning* apresentaram erros RMSE e MAE menores do que o método fundamentado no Índice Fenológico de Produção (IFP).

3.5 REFERÊNCIAS

ALFONSI, L. A. **Uso de índices fenológicos em modelos de predição de produtividade e, cafeeiro**. 2008. 104f. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba-MG.

APARECIDO, L. E. O.; ROLIM, G. S.; LAMPARELLI, R. A. C.; SOUZA, P. S.; SANTOS, E. R.. Agrometeorological models for forecasting coffee yield. **Agronomy Journal**, v. 109, n. 1, 2017.

Bernardes, T.; Moreira, M. A.; Adami, M.; Giarolla, A.; Rudorff, B. F. T.. Monitoring biennial bearing effect on coffee yield using MODIS remote sensing imagery.

Remote Sensing, v. 4, n. 9, 2012.

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. Package 'randomForest'. 2018.

CAMARGO, A. P.; CAMARGO, M. B. P. Definição e esquematização das fases fenológicas do cafeeiro arábica nas condições tropicais do Brasil. **Bragantia**, v. 60, n. 1, 2001.

CERQUEIRA, E. S. A. **Planta detalhada das áreas**: Planta planialtimétrica da Fazenda Braúna. 2004.

EPIPHANIO, J. C. N.. CBERS-4: estado atual e futuro. **XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Rio Grande do Norte**, 2009.

FAHL, J. I.; CARELLI, M. L. C. Os estudos sobre a fisiologia do cafeeiro no Instituto Agrônômico. **O Agrônômico**, v. 59, 2007.

FAHL, J. I.; CARELLI, M. L. C.; ALFONSI, E. L.; CAMARGO, M. B. P. Desenvolvimento e aplicação de metodologia para estimativa da produtividade do cafeeiro, utilizando as características fenológicas determinantes do crescimento e produção. In: **Simpósio de pesquisas dos cafés do Brasil**, 4., 2005, Londrina. **Anais...Brasília**, 2005.

FAO. **FAO Statistical Pocketbook: Coffee 2015**. Roma, 2015.

GANDHI, N.; ARMSTRONG, L. Applying machine learning techniques to predict yield of rice in humid subtropical climatic zone of India. In: **Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on**. IEEE, 2016.

GAO, B. C. NDWI – A Normalized difference water index for remote sensing of vegetation liquid water from space. **Remote sensing of environment**, v. 58, n. 3, 1996.

GOMES, A. C. S.; ROBAINA, A. D.; PEITER, M. X.; SOARES, F. C.; PARIZI, A. R. C. Modelo para estimativa da produtividade para a cultura da soja. **Ciencia rural**, v. 44, n. 1, 2014.

HUANG, C.; WYLIE, B.; YANG, L.; HOMER, C.; ZYLSTRA, G. Derivation of a tasselled cap transformation based on Landsat-8 7 at-satellite reflectance. **International Journal of Remote Sensing**, v. 23, n. 8, 2002.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote sensing of environment**, v. 25, n. 3, 1988.

KEERTHI, V., KUMAR, A. S. At-sensor solar exo-atmospheric irradiance, Rayleigh optical thickness and spectral parameters of RS-2 sensors. NRSC/SDAPSA/DQEPQCD/RS-2/TN-July2011. 2011.

KUHN, M.; WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.;

COOPER, T.; MAYER, Z.; KENKEL, B.; BENESTY, M.; LESCARBEAU, R.; ZIEM, A.; SCRUCICA, L.; TANG, Y.; CANDAN, C.; HUNT, T..Package 'caret'. 2019.

MATHER, M. P. **Computer Processing of Remotely-Sensed Images: An Introduction**. Nova York, 1999.

MEDAR, R. A.; RAJPUROHIT, V. S. A survey on machine learning techniques for crop yield prediction. **International Journal of Advance Research in Computer Science and Management Studies**, vol. 2, n. 9, 2014.

MIRANDA, J. M.; REINATO, R. A. O.; SILVA, A. B. D. Modelo matemático para predição da produtividade do cafeeiro. **Revista Brasileira de Engenharia Agrícola e Ambiental, Campina Grande**, v. 18, n.4, 2014.

NUNES, F. L.; CAMARGO, M.; FAZUOLI, L. C.; ROLIM, G. S.; PEZZOPANE, J. R. M. Modelos agrometeorológicos de estimativa da duração do estágio floração-maturação para três cultivares de café arábica. **Bragantia**, v. 69, n. 4, 2010.

OLIVEIRA, D. A. **Estimativa da produção de café por meio de índice fenológico**. 2007. 20f. Dissertação (Mestrado em Agronomia) – Universidade Federal de Lavras. Lavras-MG.

PATEL, H.; PATEL, D..A comparative study on various machine learning algorithms with special reference to crop yield prediction.**Indian Journal of Science and Technology**, v. 9, n. 22, 2016.

PETERSEN, L.. Real-Time prediction of crop yields from MODIS relative vegetation health: A continent-wide analysis of Africa. **Remote Sensing**, v. 10, n. 11, 2018.

PICOLI, M. C. A.; RUDORFF, B. F. T.; RIZZI, R.; GIAROLLA, A. Índice de vegetação do sensor MODIS na estimativa da produtividade agrícola da cana-de-açúcar. **Bragantia**, v. 68, n. 3, 2009.

PINTO, C. et al. First in-flight radiometric calibration of MUX and WFI on-board CBERS-4. **Remote Sensing**, v. 8, n. 5, 2016.

RAMESH, D.; VARDHAN, B.V. Analysis of crop yield prediction using machine learning techniques. **International Journal of Research in Engineering and Technology**, vol. 4, n. 1, 2015.

RICHARDSON, A. J.; WIEGAND, C. L. Distinguishing vegetation from soil background information.**Photogrammetric Engineering and Remote Sensing**, v. 43, n. 2, 1977.

ROCHA, H. G.; SILVA, A. B.; NOGUEIRA, D. A.; MIRANDA, J. M.; MANTOVANI, J. R. Mapeamento da produtividade do cafeeiro a partir de modelos matemáticos de predição de safra. **Coffee Science**, v. 11, n. 1, 2016.

ROSA, V. G. C.; MOREIRA, M. A.; RUDORFF, B. F. T.; ADAMI, M. Estimativa da produtividade de café com base em um modelo agrometeorológico-espectral. **Pesquisa agropecuária brasileira**, v. 45, n. 12, 2010.

SATIR, O.; BERBEROGLU, S. Crop yield prediction under soil salinity using satellite derived vegetation indices. **Field Crops Research**, v. 192, 2016.

SHIU, Y. S.; CHUANG, Y. C. Yield estimation of paddy rice based on satellite imagery: comparison of global and local regression models. **Remote Sensing**, v. 11, n. 2, 2019.

SILVA, B. A. O.; REIS, E. A. A bienalidade da cafeicultura e o resultado econômico da estocagem. **Custos e @gronegocio**. v. 9, n. 3, 2013.

SONG, R.; CHENG, T.; YAO, X.; TIAN, Y.; ZHU, Y.; CAO, W. Evaluation of Landsat-8 time series image stacks for predicting yield and yield components of winter wheat. In: **Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International**. IEEE, 2016.

TAN, B.; MASEK, J. G.; WOLFE, R.; GAO, F.; HUANG, C.; VERMOTE, E. F.; SEXTON, J. O.; EDERER, G.. Improved forest change detection with terrain illumination corrected Landsat images. **Remote Sensing of Environment**, v. 136, 2013.

TAN, B.; WOLFE, R.; MASEK, J.; GAO, F.; VERMOTE, E. F.. An illumination correction algorithm on Landsat-TM data. In: **2010 IEEE International Geoscience and Remote Sensing Symposium**. IEEE, 2010.

TEAM, R. Core et al. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, 2018.

THIAM, A.; EASTMAN, R. J. **Vegetation índices**. R.J. Eastman (Ed.), Manual of the Idrisi Selva Edition, Clarke University, Worcester, MA, USA, 2012.

4. Conclusões Gerais

Os diferentes métodos de *machine learning*, bem como distintas formas de representar as variáveis de entrada e saída dos modelos foram capazes de estimar os valores de reflectância de imagens orbitais oriundas de um satélite com base em imagens capturadas por sensores de outros satélites. Os modelos que apresentaram menores erros RMSE's na predição da reflectância de imagens orbitais de uma data distinta a data cujos dados foram usados para treinar os modelos foram: processo Gaussiano (*Gaussian Process* - `gaussprLinear`) e árvore de classificação e regressão com agregação Bootstrap (`treebag`) para as predições com base em imagens Cbers-4; regressão linear (`lm`), máquinas de vetores de suporte com Kernel linear (*support vector machine with Linear Kernel*- `svmLinear1`) e `treebag` nas predições com base em imagens Resourcesat-2; e, regressão ridge Bayesiana (*Bayesian Ridge Regression* - `bridge`) nas predições com base em imagens Landsat-8.

Foi possível estimar a produtividade do cafeeiro a partir da utilização de cinco variáveis com erros semelhantes aos erros apresentados pelos modelos com 10 e com 14 variáveis que incluíam informações espectrais, topográficas e agronômicas. A estimativa da produtividade pode ser feita com antecedência de até três meses. Todos os modelos compostos por métodos de *machine learning* apresentaram erros RMSE e MAE menores do que o método fundamentado no Índice Fenológico de Produção (IFP).