

1. Introdução

A produção de manga é uma atividade de grande expressão no cenário da fruticultura brasileira. O País figura entre os maiores produtores da fruta, e de acordo com o Anuário Brasileiro de Fruticultura (2018), além de ser autossuficiente na produção, o Brasil também é o maior exportador, com 179 mil toneladas embarcadas em 2017.

A maior área plantada de mangas no Brasil se encontra na região do Vale do São Francisco, cuja área passou por uma expansão devido ao avanço da produção e exportação, aumentando de 27,17 ha em 2017 para 30,30 ha em 2018 (Hortifruti Brasil, 2018). Dentre as variedades produzidas na região, a ‘Palmer’ tem ganhado espaço em decorrência de novos plantios e, também, dá sobre-enxertia em plantios da variedade ‘Tommy Atkins’ (Trindade, Lima, & Assis, 2015).

Tendo o mercado externo como um dos principais consumidores da manga e ao mesmo tempo um mercado exigente e competitivo, fazem-se necessários estudos sobre o processo de maturação da manga até a colheita, visto que o fruto colhido em estágio imaturo não será capaz de alcançar o nível de qualidade aceitável para o consumidor, podendo limitar também a conservação pós-colheita (Costa et al., 2017).

Os métodos tradicionais e mais utilizados na determinação da maturação e qualidade de frutos baseiam-se em processos destrutivos. Neste sentido, o desenvolvimento e estudo de técnicas alternativas que permitam a determinação de atributos de qualidade, de forma precisa e não invasiva, são de extrema importância (Goulart et al., 2013), principalmente no sentido de reduzir perdas quantitativas e qualitativas de produção.

O potencial de técnicas não destrutivas como ferramentas de avaliação e classificação de frutas vem sendo alvo de diferentes estudos. Modalidades de imagem são investigadas para a avaliação da qualidade, desde imagens do infravermelho próximo (NIR), à imagens multi e hiperespectrais, imagem de reflexão de iluminação estruturada, imagens visíveis à base de luz monocromática ou preto / branco até imagens em cores ou RGB (vermelho, verde e azul) (Li, Huang, & Zhao, 2015). Com a análise e processamento de imagens digitais é possível avaliar a mudança dos aspectos visuais dos frutos de forma objetiva, integral e representativa, assim como correlacionar com atributos físico-químicos da polpa (Nagle et al., 2016).

Estudos com mangas de diferentes cultivares utilizando imagem vêm sendo desenvol-

vidos, no entanto, diferenças entre eles e também no ambiente de cultivo podem afetar o desempenho e consistência dos índices de maturação, tanto que atualmente não há consenso sobre o índice ideal para a manga, no caso específico para variedade ‘Palmer’, que
35 tem lugar de destaque no cenário nacional e internacional. Assim, a escolha de técnicas de pré-processamento de imagens e variáveis empregadas devem ser testados para cada variedade, assim como a técnica de inferência empregada, em que, se um grande conjunto de amostras for utilizado, aumenta-se a robustez do processo de predição (Pereira et al., 2017).

40 Esses estudos ainda são incipientes no Brasil, bem como na região do Vale do São Francisco, principal pólo de produção de mangas no País, demonstrando serem ferramentas novas e promissoras para determinação de atributos de qualidade desses frutos no campo. Assim, objetivou-se com este trabalho avaliar o uso de imagens de reflectância para identificação das variáveis que devem ser extraídas para a predição de atributos de
45 qualidade de mangas ‘Palmer’ em diferentes estádios de maturação.

1.1. Trabalhos Relacionados

A seguir é percorrido sobre dez diferentes trabalhos que empregam Visão computacional para determinação de atributos de qualidade em mangas de diversas variedades.

Os autores Teoh e Syaifudin (2007) determinaram, em seus trabalhos, o peso de
50 mangas da variedade Chokanan através de uma Regressão linear. Eles inicialmente utilizaram o filtro da mediana para correção de inconsistências nas imagens, seguida da segmentação. A partir das imagens segmentadas, foi obtido o número de pixels correspondentes à manga. Utilizando esta variável de entrada em uma Regressão linear, eles obtiveram um coeficiente de correlação igual a 0,9769 e erro médio igual a 3,76%.

55 Para prever sólidos solúveis totais (SST) em mangas Chokanan, Khairunniza-Bejo e Kamarudin (2011) utilizaram o espaço de cores HSB (matiz, saturação e brilho). Os autores não realizaram pré-processamento e extraíram os valores médios de matiz, saturação e brilho a partir de uma região central na manga. Eles construíram modelos de Regressão linear para as três variáveis de entrada separadamente, visando determinar o
60 melhor canal para a predição de SST. Eles obtiveram então um coeficiente de correlação igual a -0,92 com a matiz. Para mangas possuindo SST entre as faixas 4-8, 8-13 e 13-17 °Brix, respectivamente, eles obtiveram valores de raiz do erro quadrático médio (RMSE)

iguais a 0,06, 0,02 e 0,03 °Brix.

Zheng e Lu (2012) classificaram mangas quanto ao seu estágio de escurecimento através da LS-SVM (Máquina de vetores de suporte por mínimos quadrados). As imagens dos frutos foram pré-processadas pela subtração das mesmas pelo fundo. Os autores então extraíram os valores médios do canal $L^*a^*b^*$ e as variáveis fractais *Box Counting Dimension*, *Correlation Dimension* e *Dilation Dimension*. Os autores construíram 3 classificadores: um para as variáveis de cor, outra para as variáveis fractais e outro que utilizou ambos subconjuntos. Zheng e Lu (2012) obtiveram 100% de acurácia ao empregar todas as variáveis.

Para predição de dias restantes até o apodrecimento de mangas, Nandi et al. (2014) empregaram como pré-processamento o filtro *deblurring* de Wiener e filtro da mediana. Após isto, foram extraídas as médias das intensidades RGB na manga inteira e nas regiões do cume, equador e haste, assim como a diferença destas médias e o gradiente ao longo do eixo longitudinal. As melhores variáveis de entradas foram determinadas através do algoritmo de eliminação recursiva (RFE) e SVM (Máquina de vetores de suporte). Os autores obtiveram uma acurácia média igual a 96% ao empregar, predominantemente, variáveis derivadas do canal R.

No trabalho de Vélez-Rivera et al. (2014), foi estimado o estágio de maturação de mangas da variedade Manila, utilizando como variáveis de entrada a acidez titulável, SST, firmeza, índice RPI (*Ripening Index*) e média dos pixels nos espaços de cores $L^*a^*b^*$ e HSB, totalizando dez variáveis. Nenhuma técnica de pré-processamento foi empregada; ao invés disso, os autores extraíram a faixa central ao longo do comprimento da manga para cada lado da mesma. Vélez-Rivera et al. (2014) empregaram então a técnica PCA (Análise de componentes principais) para determinar as variáveis mais significantes e MDA (*Multiple Discriminant Analysis*) para classificação. Os autores obtiveram como melhor resultado uma acurácia igual a 100% ao empregar todas as variáveis físico-químicas e as médias nos canais R, G, a^* e b^* .

Mangas das variedades Totapuri, Badami e Neelam foram utilizadas por Pandey et al. (2014), que previram a presença de doenças nas frutas e o tamanho delas. Os autores converteram as imagens para a escala de cinza, reduziram seus tamanhos e as pré-processaram com filtro da mediana e uma técnica de aguçamento. As imagens foram então segmentadas e convertidas para o espaço $L^*a^*b^*$, em que b^* foi utilizado para

95 determinar o limiar entre as mangas saudáveis e com doenças. O tamanho das mangas foi previsto através de um sistema de inferência *Fuzzy*, possuindo como entrada a área estimada e o diâmetro. Os autores obtiveram uma acurácia média igual a 93,33% para determinação da saúde das mangas e 91,41% para classificação quanto ao tamanho.

Yahaya et al. (2015) determinaram os atributos SST, acidez titulável e firmeza em 100 mangas da variedade Sala, extraíndo os valores médios no espaço RGB e utilizando comotécnica de inferência a MLR (*Multiple Linear Regression*). Nenhuma técnica de pré-processamento foi mencionada. Eles obtiveram os coeficientes de correlação 0,875, 0,814 e 0,913 para a firmeza, SST e acidez titulável respectivamente, enquanto que os valores de RMSE foram iguais a 1,392 kgf, 1,218 °Brix e 0,166 pH.

105 Os estádios de maturação de mangas Alphonso foram determinadas através dos espaços de cores RGB e HSV (matiz, saturação e valor) por Salunkhe et al. (2015). O único tratamento realizado nas imagens foi a segmentação e as variáveis extraídas foram as médias das intensidades, assim como as taxas R/G, R/B e S/H. A partir da classificação manual das mangas, foram determinados os valores das três taxas de forma 110 que as classes fossem discriminadas sem erro. A partir dos limiares obtidos para as três taxas, foi desenvolvido um algoritmo para classificação das mangas, através de comandos *if-else*. Para o modelo com apenas o canal RGB, foi obtida uma acurácia igual a 90,4%, taxa de falsos positivos igual a 2,57% e taxa de verdadeiros positivos igual a 89,77%.

Yossy et al. (2017) classificaram mangas da variedade Gincu quanto ao estágio de 115 maturação e tamanho. As imagens foram pré-processadas através das operações morfológicas de abertura e fechamento e, posteriormente, reduzidas para o tamanho 16x16 pixels. Elas foram então convertidas do espaço RGB para o HSV, de forma a determinar a cor dominante da manga. O *array* resultante, de 257 posições (256 pixels da imagem mais a cor dominante), foi utilizado como entrada em uma rede neural com função e 120 ativação do tipo sigmoide e algoritmo *back propagation*. A acurácia obtida foi igual a 94%.

Em estudos com mangas da variedade Carabao, Abarra et al. (2018) determinaram os atributos acidez titulável, açúcares totais, amido total, firmeza, acidez titulável, SST e total de açúcares reduzido, sem o emprego de pré-processamento nas imagens. As 125 variáveis extraídas consistiram nos valores médios das intensidades dos pixels nos espaços RGB, HSV e $L^*a^*b^*$, sendo utilizadas como entrada em modelos de regressão linear

para cada atributo de qualidade. Os melhores resultados alcançados foram para acidez titulável e firmeza, ao utilizar apenas o canal L*, obtendo coeficientes de correlação iguais a 0,977 e 0,968 respectivamente.

130 Na Tabela 1, é feita uma sumarização das variáveis extraídas pelos autores, atributos de qualidade estimados e variedades empregadas.

Tabela 1: Sumarização dos trabalhos relacionados.

| Autores | Atributo alvo | Variedade | Média RGB | Diferença de médias e gradiente RGB | R/G, R/B e S/H | Média HSV | Cor HSV dominante | Média L*a*b* de pixels | Número de pixels | Variáveis fractais | Diâmetro |
|-------------------------------------|---------------------------------|---|-----------|-------------------------------------|----------------|-----------|-------------------|------------------------|------------------|--------------------|----------|
| Teoh e Syaifudin (2007) | Massa | Chokanan | | | | | | | X | | |
| Khairunniza-Bejo e Kamarudin (2011) | SST | Chokanan | | | | X | | | | | |
| Zheng e Lu (2012) | Apodrecimento | Sannianmang | | | | | | X | | X | |
| Nandi et al. (2014) | Apodrecimento | Kumrapali, Amrapali, Sori, Langra, Himsagar | X | X | | | | | | | |
| Vélez-Rivera et al. (2014) | Maturação | Manila | | | | X | | X | | | |
| Pandey et al. (2014) | Tamanho e doença | Totapuri, Badami e Neelam | | | | | | X | X | | X |
| Yahaya et al. (2015) | SST, acidez titulável e firmeza | Sala | X | | | | | | | | |
| Salunkhe et al. (2015) | Maturação | Alphonso | X | | X | X | | | | | |
| Yosy et al. (2017) | Maturação | Gincu | | | | | X | | | | |
| Abarra et al. (2018) | SST, acidez titulável e firmeza | Carabao | X | | | X | | X | | | |

A partir da Tabela acima, nota-se que não há uma unanimidade ou padrão quanto às variáveis escolhidas. Para os atributos alvo SST, firmeza e acidez titulável, foram
135 utilizadas apenas variáveis baseadas em cor. A partir dos espaços de cores testados, os autores selecionaram o canal que garantiu o melhor resultado. Por outro lado, para determinação da massa, apenas o número de pixels correspondente à manga é utilizado como entrada em um modelo.

Ademais, nota-se que dentre as diversas variáveis possíveis, os autores que estimaram
140 estes atributos de qualidade limitaram-se a utilizar apenas uma variável. Assim, com este

trabalho busca-se verificar se com a utilização de mais informações visuais da manga é possível obter resultados superiores para a determinação de massa, SST, firmeza e acidez titulável.

Por fim, nota-se que em nenhum artigo a variedade 'Palmer' foi utilizada. Dessa forma, espera-se que com esse trabalho seja estabelecida a abordagem ideal para predição não destrutiva de atributos de qualidade dessa variedade, contribuindo assim para o desenvolvimento das regiões produtoras do fruto, como o Vale do São Francisco.

2. Abordagem proposta

Para determinar os aspectos visuais da manga 'Palmer' mais significativos para a predição de seus atributos de qualidade, foi conduzido um estudo investigativo quanto às variáveis utilizadas na literatura, mostradas na Tabela 1.

Para isto, a técnica utilizada na abordagem proposta foi a *Random Forest*, técnica *ensemble* que combina árvores de decisão para obtenção da variável de saída, o que a torna robusta quanto à presença de ruído nos dados e menos suscetível ao *overfitting* (Chagas et al., 2016). A relação entre as variáveis de entrada e a de saída é modelada através de um conjunto de regras de decisão, construídas por divisões binárias e recursivas dos dados de treinamento de acordo com as variáveis de entrada. Dessa forma, a *Random Forest* é capaz de modelar relacionamentos hierárquicos e não lineares (GUO et al., 2015). As regras de decisão são escolhidas de acordo com a qualidade das predições realizadas, avaliada através da métrica MSE (*Mean Squared Error*), cuja fórmula é mostrada abaixo.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Em que n é o número de amostras, y_i é o valor real da variável de saída e \hat{y}_i é o valor previsto para a variável de saída.

A partir da *Random Forest*, é possível obter a importância de cada variável de entrada do modelo, o que permite a seleção das variáveis quando aliada a técnicas como a RFE (*Recursive feature elimination*), que elimina as variáveis de forma recursiva com base em suas importâncias. Em cada iteração deste algoritmo, é treinado o modelo de predição e remove-se a variável menos relevante, até que sobre uma quantidade desejada (Menze et al., 2009), conforme mostra o diagrama da Figura ??.

170 *Random Forest* forneceu bons resultados nos trabalhos de Granitto et al. (2006) e Zhou et al. (2014).

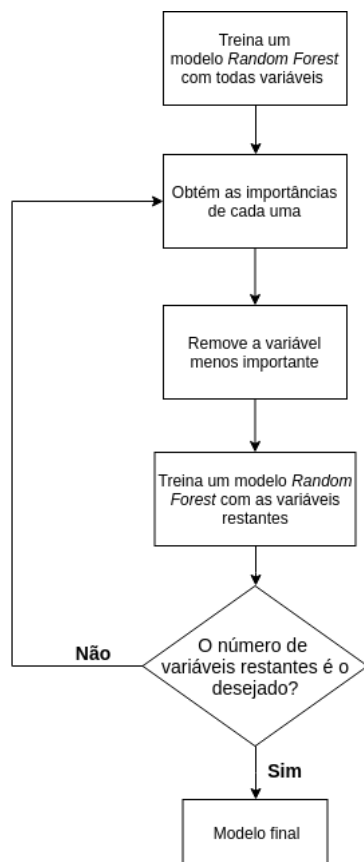


Figura 1: Diagrama de representação do modelo *Random Forest* aliado à RFE.

Assim, foram construídos modelos *Random Forest* para cada subconjunto de variáveis identificado nos trabalhos relacionados, que foram codificados conforme a Tabela 2. Ademais, foi utilizado o subconjunto que engloba todas as variáveis, a partir do qual foi realizada a seleção de variáveis para redução da dimensionalidade e complexidade do modelo.

Tabela 2: Subconjuntos de variáveis obtidos a partir da literatura.

| Id | Nº | Variáveis |
|----|----|---|
| G1 | 1 | Média das intensidades dos pixels no canal L* |
| G2 | 1 | Cor HSV dominante |

| | | |
|-----|----|---|
| G3 | 20 | Média das intensidades RGB na manga inteira, nas regiões apex, equatorial e cume, diferença das médias e gradiente longitudinal |
| G4 | 1 | Número de pixels correspondente à manga |
| G5 | 3 | Média das intensidades dos pixels no espaço RGB |
| G6 | 6 | Média das intensidades dos pixels no espaço L*a*b* e variáveis fractais (box counting dimension, dilation dimension e correlation dimension) |
| G7 | 1 | Média das intensidades dos pixels no canal Hue |
| G8 | 6 | Média das intensidades dos pixels nos espaços HSV e L*a*b* |
| G9 | 9 | Média das intensidades dos pixels nos espaços RGB e HSV e taxas R/G, R/B e S/H. |
| G10 | 3 | Média das intensidades dos pixels no canal b*, número de pixels e diâmetro |

Os modelos foram construídos inicialmente para predição da idade das mangas, que variaram entre 35 dias após a floração (DAF) e 20 dias após a colheita (DAC). A partir da determinação do melhor subconjunto de variáveis, foram determinados os atributos de massa, SST, firmeza e acidez titulável. O experimento foi assim conduzido devido à relação entre estes atributos de qualidade e a maturação da fruta, que sofre trans-
180 formações físicas e químicas durante o processo, resultando em modificações na textura, pigmentação e sabor (Tucker, 1993). Conforme a manga amadurece, são observados o aumento do teor de sólidos solúveis totais e diminuição da acidez titulável e firmeza (Mattoo et al., 1975 e Wills et al., 1981).

Na Figura 2, é mostrado o diagrama de fluxo do experimento, em que são especificadas
185 as etapas para predição dos atributos de qualidade em mangas 'Palmer'.

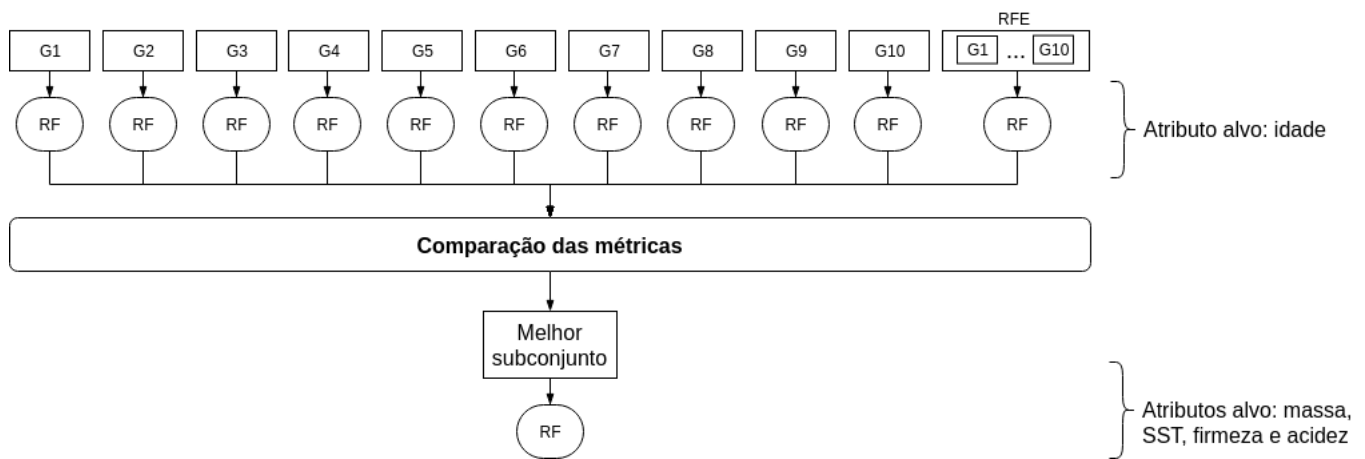


Figura 2: Diagrama de fluxo do experimento.

3. Material e Métodos

3.1. Amostras

Foram utilizadas mangas cv. ‘Palmer’ coletadas manualmente em pomar comercial da Fazenda Special Fruit Importação e Exportação Ltda., localizada no município de Petrolina–Pernambuco, região de clima do tipo BSwH (semiárido, tipo estepe, muito quente, com estação chuvosa no verão), segundo classificação de Koppen, que fica a 9°18’13,5”S e 40°40’04,7”O, com altitude de aproximadamente 380 m.

A Fazenda tem 114 hectares plantados com manga ‘Palmer’. O lote em estudo tem 3,47 há, com espaçamento da cultura de 6 X 4 m e é utilizado o sistema de irrigação de microaspersão, com turno de rega diário e com lâmina ajustada ao longo do ciclo. As mangueiras receberam todos os tratos culturais de acordo com as exigências da cultura.

Foram selecionadas trinta plantas, distribuídas em cinco fileiras de plantio de um lote do pomar. Foram coletados 600 frutos no total, oitenta em cada fase: 35, 50, 65, 80, 95, 110, 125, 140, 165 e 180 dias após a floração (DAF), ponto de colheita comercial adotado pela Fazenda. A colheita foi realizada no período da manhã, utilizando-se uma tesoura de poda para o corte do pedúnculo.

Após a coleta, os frutos foram transportados cuidadosamente até o Laboratório de Armazenamento de Produtos Agrícolas (LAPA) da Universidade Federal do Vale do São Francisco, campus Juazeiro-BA, onde foram lavados em água corrente, um a um, e imersos em solução de 150 mg de cloro por litro de água por 15 minutos, com posterior enxague para remoção do excesso de cloro e secagem em temperatura ambiente. Em

seguida, foi determinada a massa dos frutos com auxílio de balança semi-analítica com precisão de 0.01 g.

210 A firmeza do fruto foi determinada com o auxílio de um penetrômetro digital modelo PTR 300, com ponteira de 6 mm de diâmetro. Foi realizada uma leitura por fruto, na porção equatorial, e o resultado foi expresso em Newtons (N). Os sólidos solúveis totais (SST) foram determinados de forma destrutiva em filtrado da polpa centrifugada, utilizando um refratômetro digital (Hanna – HI 96804), sendo os resultados expressos em 215 °Brix (AOAC, 1997). Por fim, a acidez titulável foi determinada através de titulação com solução de hidróxido de sódio (0.1 M NaOH) com 1% de fenolftaleína como indicador, também de acordo com a metodologia da AOAC (1997).

Os valores de referência supracitados foram determinados após o processo de aquisição das imagens, realizada em parceria com o Laboratório de Energia na Agricultura (LENA) 220 da mesma Universidade. O sistema de aquisição de imagens de reflectância foi constituído por uma câmera fotográfica Canon T5i, caixa de interior preto fosco, fonte de alimentação ajustável e uma caixa de controle de acendimento de LEDs, representado na Figura 3. O sistema de iluminação era constituído por 3 LEDs Solderless XPE2 de 3W da CREE, branco frio 5000K a 8300K, dispostos a uma distância angular de 120° entre eles.

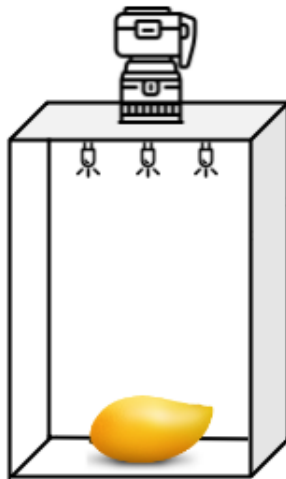


Figura 3: Esquema de aquisição das imagens.

225 Para o processo de aquisição das imagens foi obtida uma imagem para cada lado de cada fruto (considerando a posição de repouso), através da câmera ajustada com foco manual, ISO-100; tempo de exposição 1/2s; F/5.6; distância focal de 48mm.

3.2. Pré-processamento

Como as imagens obtidas possuíam ruído e informações irrelevantes, foi necessário realizar o pré-processamento das mesmas. Foram testadas diferentes configurações de algoritmos, de forma a remover a maior quantidade possível de ruído sem perda de informações relevantes da imagem. Apesar das fotos terem sido tiradas em um ambiente controlado, as imagens resultantes variaram quanto ao ruído nelas contido, devido ao acúmulo de sujeira na câmara com o passar das semanas do experimento. Sendo assim, as configurações das técnicas variaram para cada imagem.

A primeira técnica aplicada foi o filtro da mediana, visando a remoção de manchas contidas nas imagens. O tamanho de janela que garantiu uma melhor remoção de ruído na maioria das imagens foi de 11x11 pixels. Após isso, as imagens foram segmentadas através do algoritmo de Otsu, de forma que as mangas fossem isoladas do fundo. Com isso, notou-se nas imagens a presença de pequenos pontos que não foram removidos pelo filtro da mediana. Apesar de a remoção dos mesmos ter sido possível com uma segunda filtragem pela mediana, notou-se uma perda de detalhes na manga. Sendo assim, optou-se por empregar a operação de abertura, operação morfológica através da qual pequenos pontos de uma imagem podem ser removidos.

A utilização destas técnicas não garantiu uma remoção completa das sombras contidas na imagem. Assim, utilizou-se a limiarização simples para este fim, visto que a intensidade dos pixels correspondentes às sombras era, em geral, visivelmente menor que a intensidade dos pixels das mangas. Como a limiarização simples resultou na remoção de partes da manga, além das sombras delas, empregou-se a operação morfológica de fechamento para o preenchimento das mangas.

Por fim, foram traçados os contornos das mangas, visando a remoção de partes da imagem que não continham a fruta. A Figura 4 mostra uma imagem original e pré-processada.

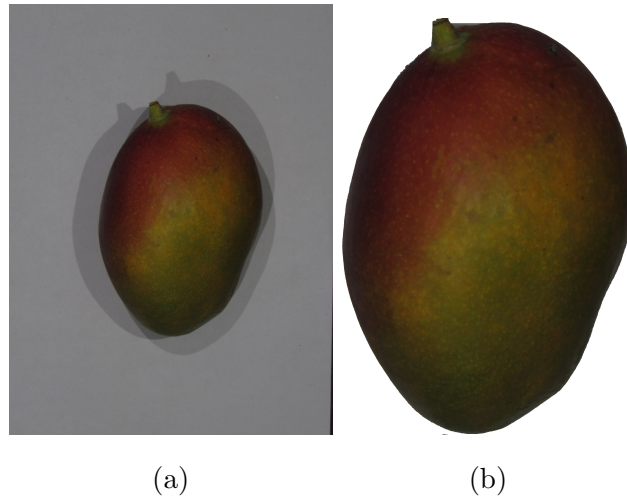


Figura 4: Efeito do pré-processamento. (a) Imagem original (b) Imagem pré-processada.

Posteriormente, as variáveis mencionadas na Tabela 2 foram extraídas para cada
255 imagem. Todos os algoritmos de pré-processamento de imagens, assim como a extração
de variáveis, foram realizados através da linguagem de programação Python (versão
3.6.7), com o auxílio da biblioteca OpenCV (versão 3.4.4).

3.3. Método e métricas de avaliação de desempenho

Antes da construção dos modelos foi realizada uma reamostragem das imagens, devido
260 ao desbalanceamento presente na base de dados. Como havia 600 imagens de mangas
tiradas antes da colheita e apenas 240 imagens tiradas após ela, foi feita uma repetição
aleatória das imagens deste período até que fosse obtida uma quantidade igual de imagens
nos dois períodos.

Para a avaliação da capacidade preditiva dos modelos, foi utilizada a estratégia de
265 validação cruzada *5-fold* (*5-fold cross validation*). Ela é empregada para assegurar a
robustez do modelo, através da divisão do conjunto de dados em 5 subconjuntos disjun-
tos, com uma alocação das amostras para o conjunto de treinamento ou teste (Zhang
Yang, 2015). Assim, um dos subconjuntos foi utilizado como teste e os 4 demais para
o treinamento, de forma que o modelo realizasse a predição para dados desconhecidos.
270 Este procedimento foi repetido 5 vezes, alterando os subconjuntos a cada vez.

As predições resultantes foram avaliadas através de métricas, tais como o coeficiente
de correlação e raiz do erro quadrático médio (RMSE), empregadas nos artigos em que a
regressão foi realizada. Estes indicadores medem, respectivamente, o grau de dependência

entre as variáveis de entrada e saída e a magnitude média dos erros estimados (Alves;
275 Vecchia, 2011), conforme as equações abaixo:

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

Em que x_i é o valor da variável de entrada, \bar{x} é a média dos valores de x , y_i é o valor real da variável de saída, \bar{y} é a média dos valores de y , n é o número de amostras e \hat{y}_i é o valor previsto para a variável de saída.

3.4. Abordagem comparada

Na Tabela 3, são mostradas as variáveis de entrada utilizadas na literatura para
280 predição de massa, SST, firmeza ou acidez titulável, assim como a técnica de inferência utilizada. Nota-se que para todos os trabalhos foi utilizada a Regressão linear para predição do atributo alvo. Entretanto, esta técnica assume um relacionamento linear entre as variáveis de entrada e de saída, o que, apesar de ter fornecido bons resultados
285 para as variedades empregadas pelos autores, pode não ser o caso para a 'Palmer'.

Na predição de cada um dos atributos de qualidade, foi comparada a abordagem empregada na literatura e o melhor subconjunto obtido pela *Random Forest*. As mesmas variáveis utilizadas pelos autores serão utilizadas como entrada em uma Regressão linear e os resultados obtidos serão comparados diretamente aos resultados obtidos pelo melhor
290 modelo *Random Forest*.

Tabela 3: Atributo de qualidade, variáveis de entrada e técnica de inferência empregadas na literatura.

| Atributo de qualidade | Variáveis de entrada | Técnica de inferência |
|-----------------------|----------------------|-----------------------|
| Massa | Número e pixels | Regressão linear |
| SST | Valor médio da matiz | Regressão linear |
| Firmeza | Valor médio de L* | Regressão linear |
| acidez titulável | Valor médio de L* | Regressão linear |

A partir da comparação dos resultados, será possível determinar a melhor abordagem para predição de atributos de qualidade em mangas 'Palmer'.

3.5. Análise estatística

4. Resultados e Discussão

295 Após a construção de modelos de determinação de idade das mangas para cada subconjunto de variáveis da Tabela 2, foram obtidas as métricas coeficiente de correlação e RMSE (Raiz do erro quadrático médio). Nos gráficos da Figura 5, são mostrados os resultados obtidos para a *Random Forest*.

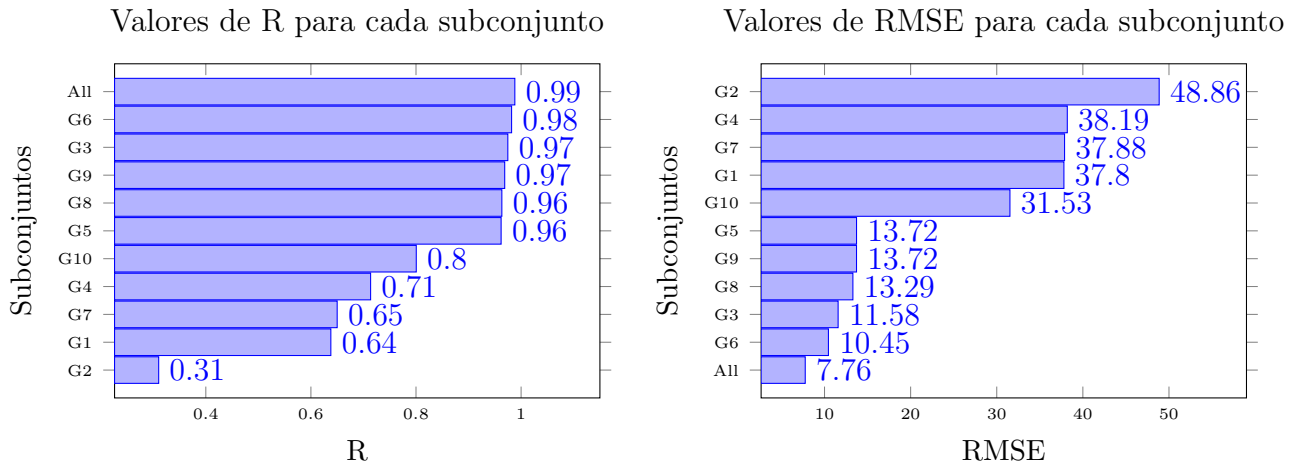


Figura 5: Métricas obtidas na determinação da idade para os subconjuntos de variáveis (a) R (b) RMSE.

Nota-se que os melhores resultados foram obtidos para a *Random Forest* quando feita a seleção automática de variáveis. Através da técnica RFE (*Random Feature Elimination*), foi obtido que as variáveis mais significantes foram as taxas R/B, R/G e S/H, média do canal B na haste, cume e equador, média do canal R na haste e cume, médias dos canais R, B, G, H e *a na manga inteira, diferença entre R e B e R e G na manga inteira, diferença entre R do cume e do equador e do cume e haste, diferença entre B do cume e haste, dimensão de correlação, área e diâmetro. Com mais informações extraídas da manga, torna-se mais provável um melhor resultado para predição da idade. Optou-se então por utilizar estas variáveis para a predição de massa, SST (Sólidos solúveis totais), firmeza e acidez titulável através da técnica *ensemble*.

4.1. Estatística descritiva

310 Após a determinação do melhor modelo para predição do tempo, foram treinadas *Random Forests*, com as vinte variáveis mais significantes, para predição de massa, SST,

315 firmeza e acidez titulável. Os modelos resultantes foram comparados aos melhores modelos da literatura. Na Tabela 4, é mostrada a estatística descritiva dos atributos de qualidade, contendo a quantidade, média, valor mínimo e máximo, amplitude, desvio padrão e variância dos valores de referência.

Tabela 4: Estatística descritiva dos valores reais de massa, SST, firmeza e acidez titulável.

| Atributos | Amostras | Média | Mín | Máx | Amp | SD | Var |
|------------------|----------|--------|------|--------|--------|--------|----------|
| Massa | 1200 | 436,51 | 26,7 | 757,35 | 730,65 | 136,22 | 18557,30 |
| SST | 1200 | 9,83 | 3,8 | 19,7 | 15,9 | 4,49 | 20,24 |
| Firmeza | 1200 | 74,048 | 2,95 | 181,4 | 178,45 | 60,80 | 3697,50 |
| Acidez titulável | 1200 | 0,69 | 0,03 | 9,97 | 9,93 | 0,64 | 0,42 |

Percebe-se uma grande variabilidade nos dados, o que garantiu uma maior robustez aos modelos construídos, que puderam realizar previsões para mangas 'Palmer' em diferentes estágios de maturação. Nas seções posteriores são discutidos os resultados da previsão de massa, SST, firmeza e acidez titulável.

320 4.2. Massa

Com a previsão de massa através da abordagem proposta e da metodologia dos autores Teoh e Syaifudin (2007), foram obtidos os valores de R e RMSE por *fold*. Na Figura 6, é feita uma comparação das métricas obtidas.

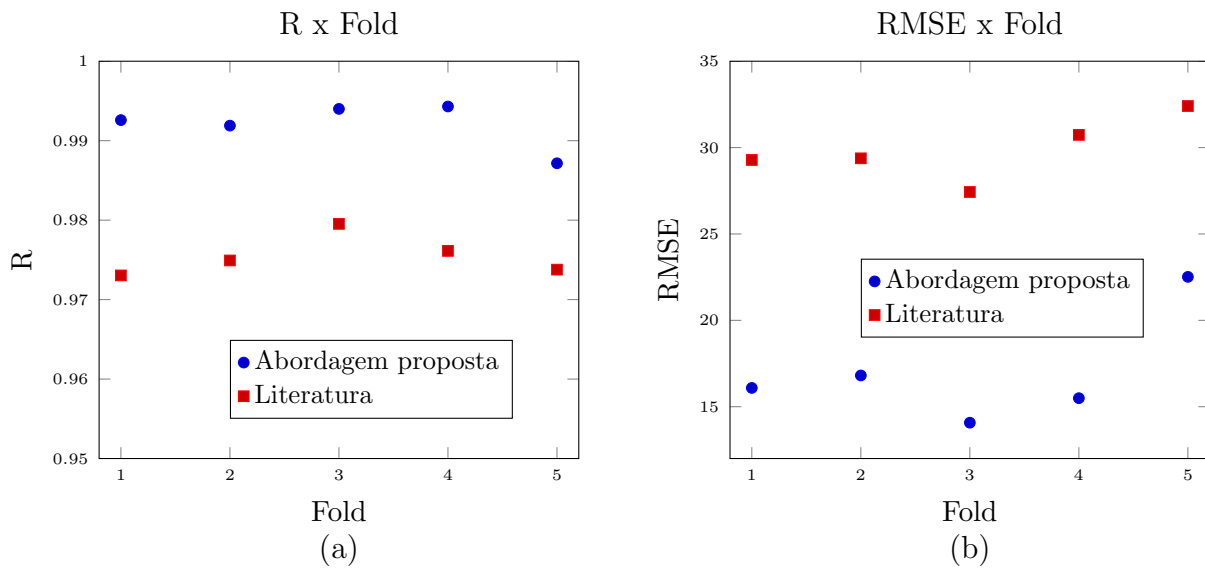


Figura 6: Métricas obtidas na previsão de massa de frutos da manga 'Palmer' (a) R (b) RMSE.

Nota-se que a *Random Forest* foi capaz de prever massa em mangas de forma mais
 325 precisa que a Regressão linear. O coeficiente de correlação médio para a técnica *ensemble*
 foi igual a 0,9920 e o erro médio igual a 2,98%. Os autores Teoh e Syaifudin (2007)
 obtiveram, por outro lado, um coeficiente de correlação igual a 0,9769 e erro médio igual
 a 3,76% para mangas da variedade Chokanan. Na Figura 7, são mostrados os gráficos
 de dispersão para o melhor modelo encontrado na literatura e para o modelo *Random*
 330 *Forest*. Nota-se um melhor ajuste dos dados à reta para a técnica *ensemble*.

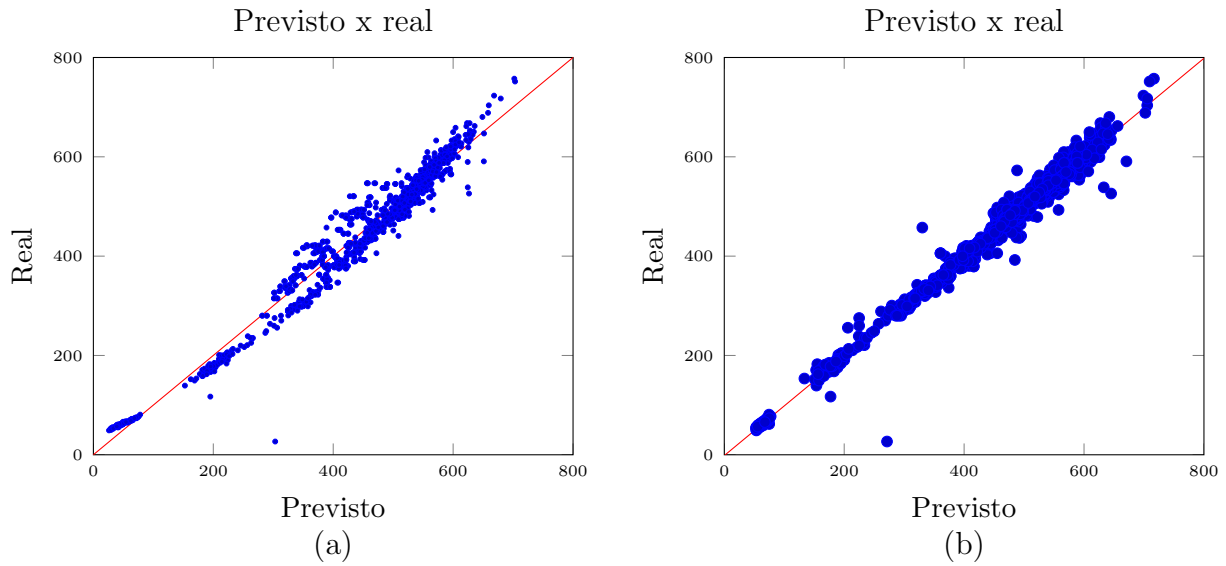


Figura 7: Gráficos de dispersão para massa (a) Modelo da literatura (b) Abordagem proposta.

Na Tabela 5, é feito um resumo dos testes de hipótese para massa.

Tabela 5: Resumo dos testes de hipótese para massa.

| Métrica | μ_1 | μ_2 | $\mu_1 - \mu_2$ | p -value |
|---------|---------|---------|-----------------|------------|
| RMSE | 16.9961 | 29.8469 | -12.8507 | |
| R | 0.9919 | 0.9754 | 0.0165 | |

4.3. Sólidos solúveis totais (SST)

Ao empregar a metodologia de Khairunniza-Bejo e Kamarudin (2011) para predição
 de SST, foi obtido um coeficiente de correlação médio inferior a 0.1 e RMSE médio
 335 igual a 4.48, enquanto que os autores alcançaram valores iguais a -0,92 e 0,033 °Brix
 respectivamente para mangas da variedade Chokanan.

Nota-se que para a variedade 'Palmer', um modelo linear que emprega apenas uma variável, a matiz, é insuficiente para a determinação precisa de SST. Por outro lado, ao empregar a abordagem proposta, foram obtidos valores médios de R e RMSE iguais a 0,9797 e 0,8946 °Brix. Na Figura 8, são mostradas as métricas por *fold* para a Regressão linear e *Random Forest*.

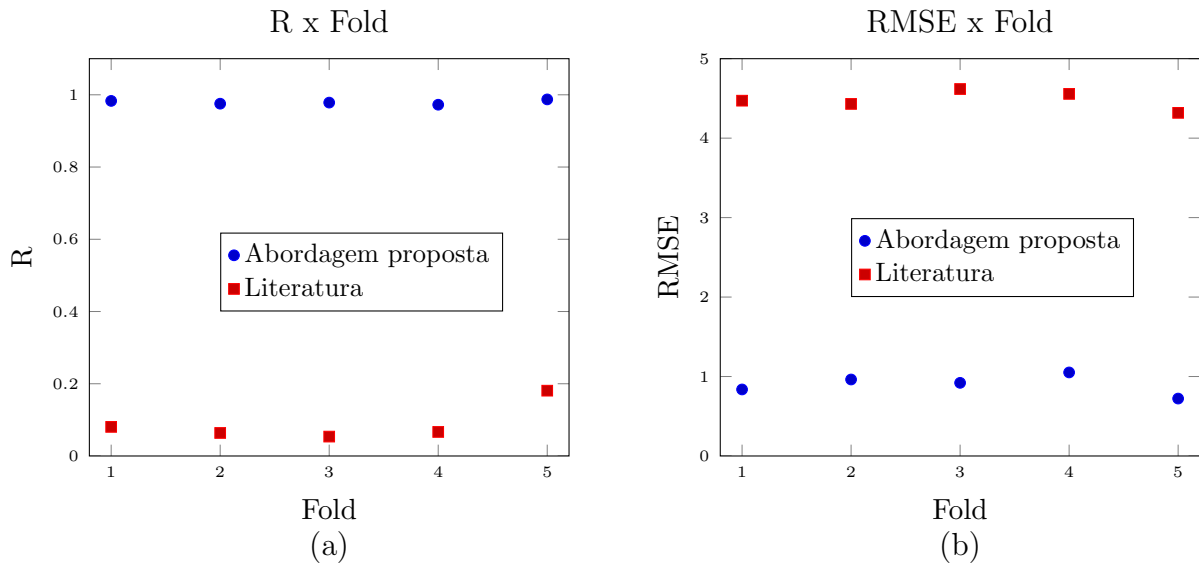


Figura 8: Métricas obtidas na predição de SST de frutos da manga 'Palmer' (a) R (b) RMSE.

Assim, o valor de R obtido no presente estudo foi superior aos encontrados na literatura. Por outro lado, o RMSE ainda foi maior que o obtido por Khairunniza-Bejo e Kamarudin (2011).

Na Figura 9, são mostrados os gráficos de dispersão para a Regressão linear com a matiz de entrada e para o modelo *Random Forest* com as melhores variáveis.

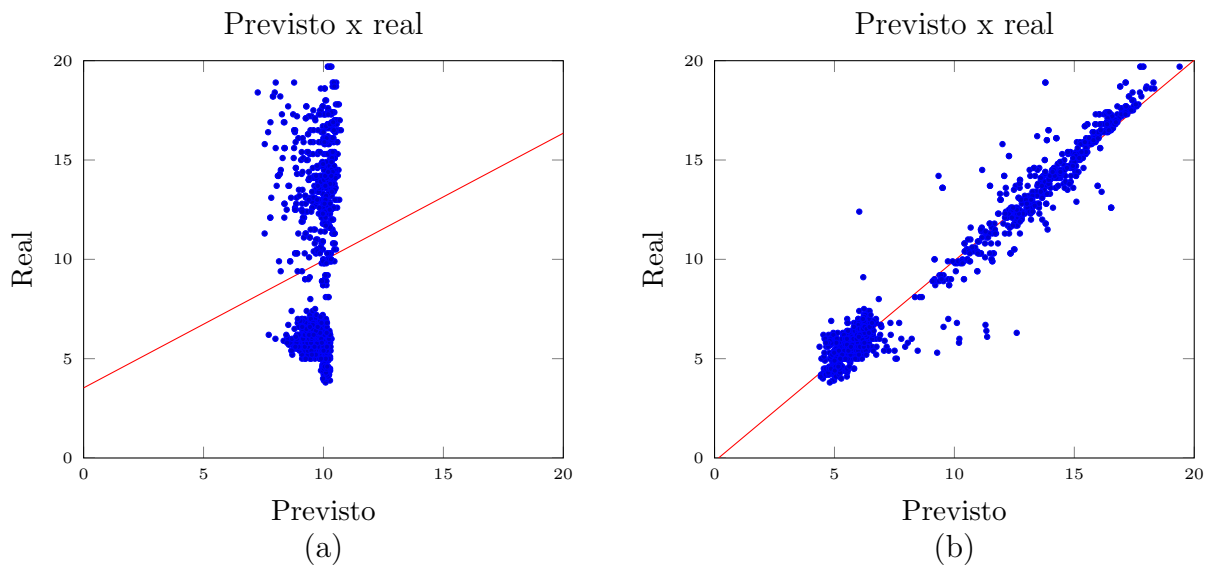


Figura 9: Gráficos de dispersão para SST (a) Modelo da literatura (b) Abordagem proposta.

Mais uma vez foi obtido um melhor ajuste dos dados para a técnica *ensemble*. Por possuir mais informações da manga como entrada do modelo, o relacionamento com o SST foi modelado de forma mais precisa.

Na Tabela 6, é feito um resumo dos testes de hipótese para SST.

Tabela 6: Resumo dos testes de hipótese para SST.

| Métrica | $\mu1$ | $\mu2$ | $\mu1 - \mu2$ | p -value |
|---------|--------|--------|---------------|------------|
| RMSE | 0.8988 | 4.4790 | -3.5802 | |
| R | 0.9793 | 0.0890 | 0.8902 | |

4.4. Firmeza

Ao empregar o valor médio dos pixels no canal L^* como entrada em uma Regressão linear, foi obtido um valor médio de R igual a 0,1860, muito menor que o obtido por Abarra et al. (2018), que obtiveram um coeficiente igual a 0,968. Assim, nota-se que essa variável não varia linearmente para a variedade 'Palmer', diferentemente da variedade empregada pelos autores, a Carabao.

Por outro lado, através do modelo *Random Forest*, que emprega as vinte variáveis mais significantes, foi alcançado um coeficiente de correlação médio igual a 0,9674, ainda menor que o obtido pelos autores, mas bastante maior que o obtido anteriormente. Os valores de R e RMSE por *fold* são mostrados na Figura 10.

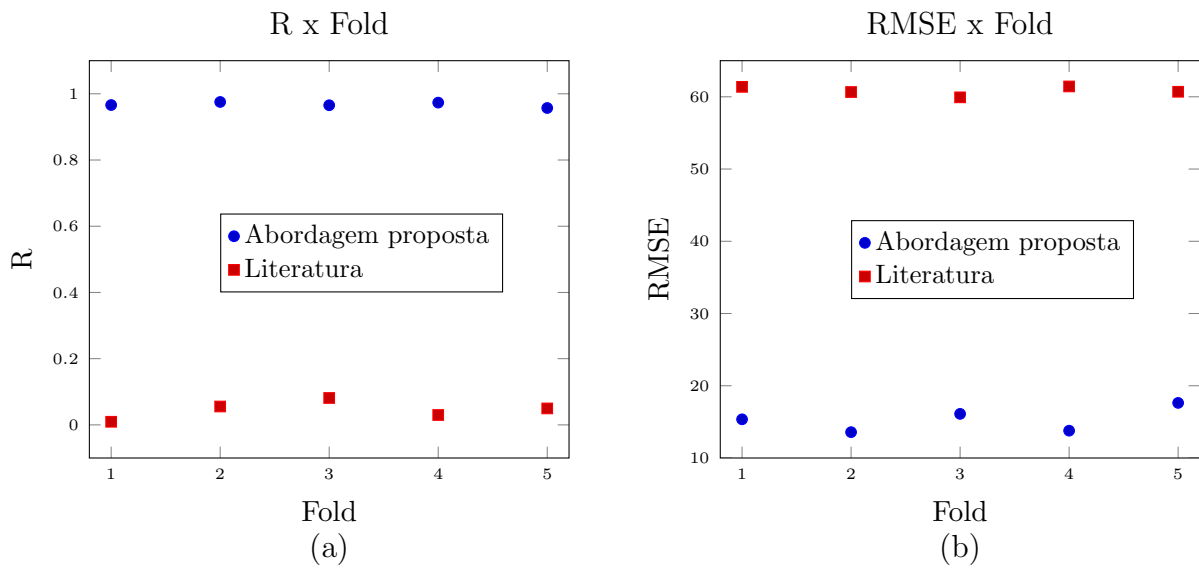


Figura 10: Métricas obtidas na predição de firmeza de frutos da manga 'Palmer' (a) R (b) RMSE.

Na Figura 11, são mostrados os gráficos de dispersão para os dois modelos de predição de firmeza. Nota-se um ajuste muito pobre para o modelo linear, enquanto que no modelo não linear os dados concentram-se mais ao redor da reta.

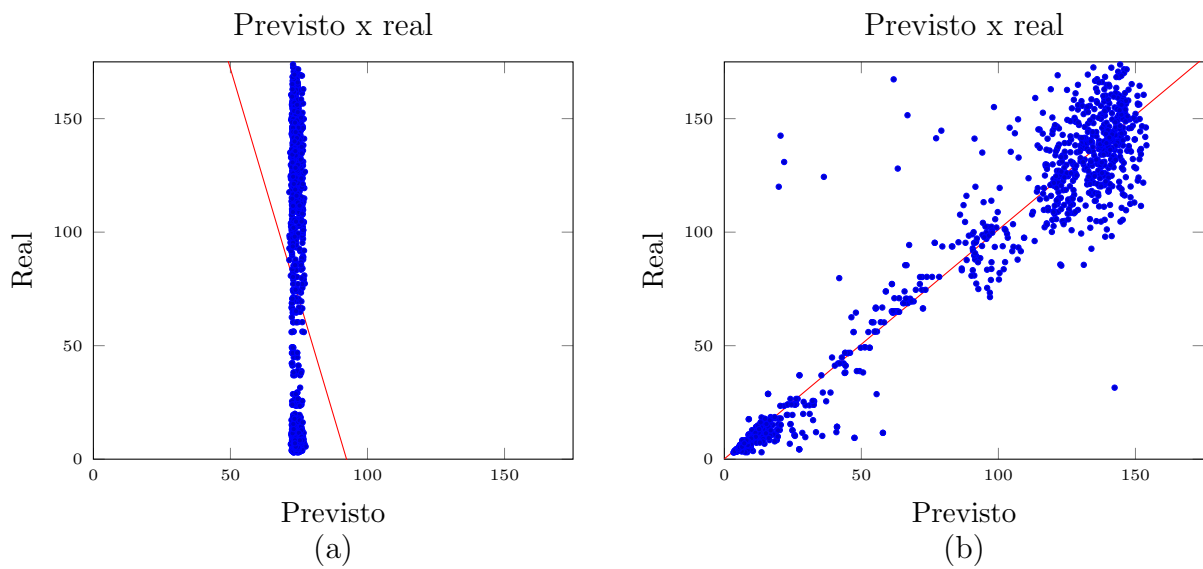


Figura 11: Gráficos de dispersão para firmeza (a) Modelo da literatura (b) Abordagem proposta.

Na Tabela 7, é feito um resumo dos testes de hipótese para firmeza.

Tabela 7: Resumo dos testes de hipótese para firmeza.

| Métrica | $\mu 1$ | $\mu 2$ | $\mu 1 - \mu 2$ | p -value |
|---------|---------|---------|-----------------|------------|
| RMSE | 15.2894 | 60.8198 | -45.5303 | |
| R | 0.9674 | 0.04516 | 0.9222 | |

365 4.5. Acidez titulável

Da mesma forma que para a firmeza, foi utilizada a variável L^* como entrada em um modelo linear. Enquanto que os autores Abarra et al. (2018) alcançaram um coeficiente de correlação igual a 0,977 para a variedade Carabao, foi obtido um valor de R igual a 0,3882 para a 'Palmer'. Neste caso, também, o atributo Acidez titulável não possui uma
370 relação linear com a variável L^* .

Ao empregar a *Random Forest*, foram obtidos resultados significativamente maiores, com um valor médio de R igual a 0,9077, mas ainda menor que o obtido pelos autores.

As métricas para ambos modelos são mostrados na Figura 12.

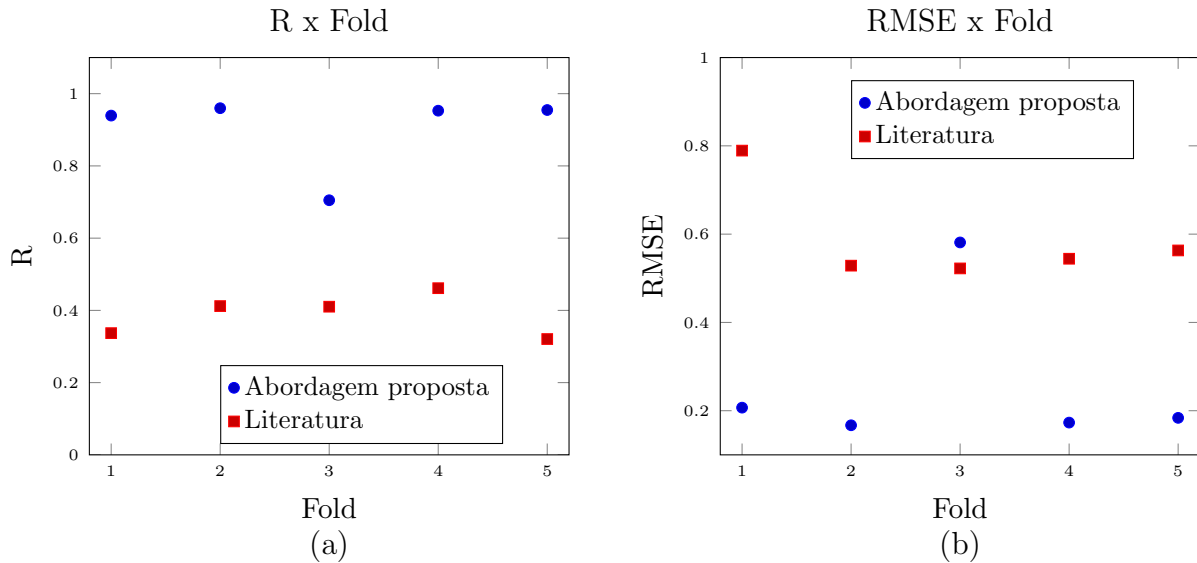


Figura 12: Métricas obtidas na predição de acidez titulável de frutos da manga 'Palmer' (a) R (b) RMSE.

Na Figura 13, são mostrados os gráficos de dispersão para os dois modelos de predição de acidez titulável. O modelo proposto é claramente superior ao sugerido para a literatura, mas ainda sim inferior ao obtido pelos autores. Nota-se que, assim como para a firmeza, a relação da acidez titulável com o valor médio de L^* não é linear. Isto pode
375

ser explicado pelo fato de que as mangas utilizadas pelos autores eram da variedade Carabao, e não 'Palmer'.

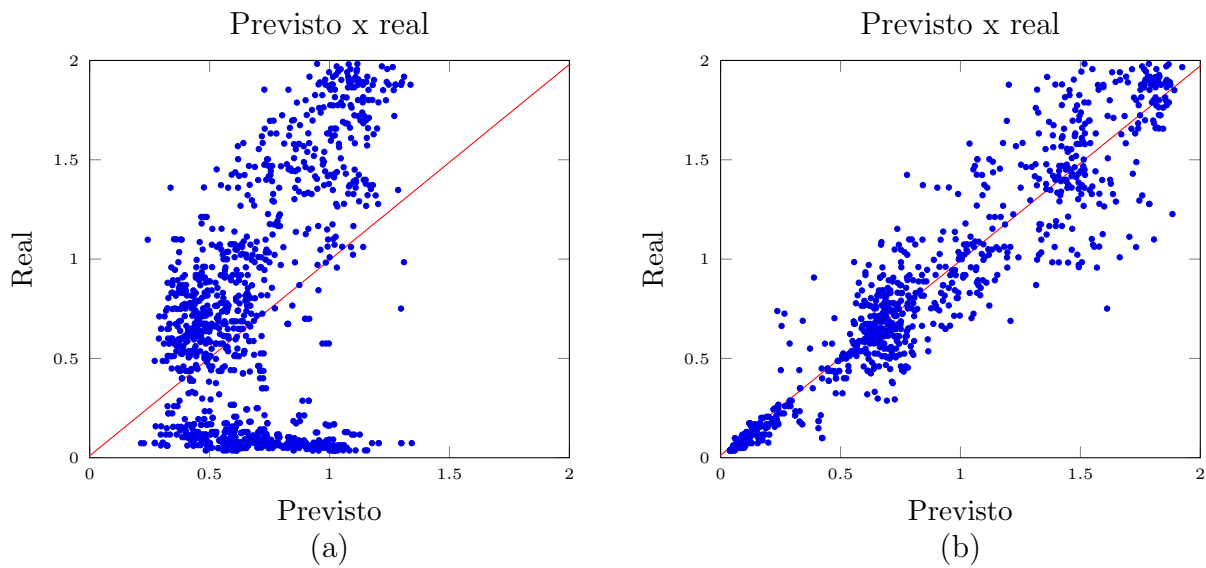


Figura 13: Gráficos de dispersão para acidez titulável (a) Modelo da literatura (b) Abordagem proposta.

380 Na Tabela 8, é feito um resumo dos testes de hipótese para acidez titulável.

Tabela 8: Resumo dos testes de hipótese para acidez titulável.

| Métrica | μ_1 | μ_2 | $\mu_1 - \mu_2$ | p -value |
|---------|---------|---------|-----------------|------------|
| RMSE | 0.2624 | 0.5894 | -0.3270 | |
| R | 0.9023 | 0.3882 | 0.5140 | |

5. Referências

Abarra, Maja Sierhine J., et al. "Determination of Fruit Ripeness Degree of 'Carabao'Mango (*Mangifera indica* L.) using Digital Photometry." *Philippine Journal of Science* 147.2 (2018): 249-253.

385

Anuário brasileiro da fruticultura 2018 / Benno Bernardo Kist... [et al.]. – Santa Cruz do Sul : Editora Gazeta Santa Cruz, 2018. 88 p. Disponível em:
[http://www.editoragazeta.com.br/sitewp/wp-content/uploads/2018/04/s FRUTICULTURA_2018_dupla.pdf](http://www.editoragazeta.com.br/sitewp/wp-content/uploads/2018/04/s_FRUTICULTURA_2018_dupla.pdf).

390

Chagas, César D. S, et al. "Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions." *Catena* 139 (2016): 232-240.

395

Costa, J. D. S., Almeida, F. D. C., Figueiredo Neto, A., Cavalcante, I. H. L. Physical and mechanical parameters correlated to the ripening of mangoes (*Mangifera indica* L.) cv. 'Tommy Atkins', *Acta Agronómica*, 66 (2) (2017), pp. 186-192.
<http://dx.doi.org/10.15446/acag.v66n2.54757>.

400

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

405

Goulart, C., Loy, F. S., Galarça, S. P., Giovanaz, M. A., Malgarim, M. B., Fachinello, J. C. Evolução do índice DA e coloração da epiderme de mangas da cultivar Tommy Atkins. *Revista Iberoamericana de Tecnología Postcosecha*, 14 (2013), pp. 8-13. Disponível em:
<http://www.redalyc.org/html/813/81327871003/index.html>.

Granitto, Pablo M. et al. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, v. 83, n. 2, p. 83-90, 2006.

410

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping

of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. *Geoderma* 237–238, 49–59.

415 Hortifruti Brasil. Anuário 2017 – 2018. Disponível em: www.hfbrasil.org.br/br/revista/acessar/completo/anuario-2017-2018.aspx. Acesso em: 14 fev. 2019.

Khairunniza-Bejo, Siti, and Syahidah Kamarudin. "Chokanan mango sweetness determination using hsb color space." 2011 Third International Conference on Computational
420 Intelligence, Modelling Simulation. IEEE, 2011.

Li, J. B., Huang, W. Q., Zhao, C. J. Machine vision technology for detecting the external defects of fruits. *Imaging Science Journal*, 63 (5) (2015), pp. 241-251. doi: 10.1179/1743131X14Y.0000000088.

425 Lima Alves, Elis Dener, and Francisco Arthur Silva Vecchia. "Análise de diferentes métodos de interpolação para a precipitação pluvial no Estado de Goiás." *Acta Scientiarum. Human and Social Sciences* 33.2 (2011).

430 Mattoo, A. K.; Modi, V. V. Palmitic acid activation of peroxidase and its possible significance in mango ripening. *Biochimica et Biophysica Acta (BBA)-Enzymology*, v. 397, n. 2, p. 318-330, 1975.

Menze, Bjoern H. et al. A comparison of random forest and its Gini importance with
435 standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, v. 10, n. 1, p. 213, 2009.

Nagle, M., Intani, K., Romano, G., Mahayothee, B., Sardsud, V., Müllher, J. Determination of surface color of 'all yellow' mango cultivars using computer vision. *International
440 Journal of Agricultural and Biological Engineering*, 9 (1) (2016), pp. 42-50. Disponível em: <https://ijabe.org/index.php/ijabe/article/view/1861/pdf>.

Nandi, Chandra Sekhar, Bipan Tudu, and Chiranjib Koley. "A machine vision-based

maturity prediction system for sorting of harvested mangoes." IEEE Transactions on Instrumentation and measurement 63.7 (2014): 1722-1730.

AOAC, 1997. Official Methods of Analysis of the Association of Official Analytical Chemists, sixteenth ed. Patricia Cuniff, Arlington.

Pandey, Rashmi, Nikunj Gamit, and Sapan Naik. "Non-destructive quality grading of mango (*Mangifera Indica* L) based on CIELab colour model and size." 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. IEEE, 2014.

Salunkhe, Rahul Pralhad, and Aniket Anil Patil. "Image processing for mango ripening stage detection: RGB and HSV method." 2015 Third International Conference on Image Information Processing (ICIIP). IEEE, 2015.

Teoh, C. C., and AR Mohd Syaifudin. "Image processing and analysis techniques for estimating weight of Chokanan mangoes." Journal of Tropical Agriculture and Food Science 35.1 (2007): 183.

Trindade, D. C. G., Lima, M. A. C., Assis, J. S. Ação do 1-metilciclopropeno na conservação pós-colheita de manga 'Palmer' em diferentes estádios de maturação. Pesquisa Agropecuária Brasileira, 50 (9) (2015), pp. 753-762. <http://dx.doi.org/10.1590/S0100-204X2015000900003>.

Tucker, G. A. Introducion. In: SEYMOUR, G. B. et al. Biochemistry of fruit ripening. London: Chapman Hall. 1993. Cap. 1, 255-266 p.

Vélez-Rivera, Nayeli, et al. "Computer vision system applied to classification of "Manila" mangoes during ripening process." Food and bioprocess technology 7.4 (2014): 1183-1194.

Wills, R. B. H. et al. Temperature. In: Postharvest physiology, handling of fruits and vegetables. Austrália: N. S. W. U. Press. 1981. 39-51 p.

Yahaya, Ommi Kalsom Mardziah, et al. "Determining Sala mango qualities with the use of RGB images captured by a mobile phone camera." AIP Conference Proceedings. Vol. 1657. No. 1. AIP Publishing, 2015.

480

Yossy, Emny Harna, et al. "Mango Fruit Sortation System using Neural Network and Computer Vision." Procedia computer science 116 (2017): 596-603.

Zhang, Yongli; Yang, Yuhong. Cross-validation for selecting a model selection procedure. Journal of Econometrics, v. 187, n. 1, p. 95-112, 2015.

485

Zheng, Hong, and Hongfei Lu. "A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.)." Computers and Electronics in Agriculture 83 (2012): 47-51.

490

Zhou, Qifeng et al. Structure damage detection based on random forest recursive feature elimination. Mechanical Systems and Signal Processing, v. 46, n. 1, p. 82-90, 2014.