UNIVERSIDADE FEDERAL DE VIÇOSA

Aplicação de técnicas estatísticas e do sensoriamento remoto no melhoramento genético do tomateiro para produtividade, qualidade e resistência à requeima

Felipe de Oliveira Dias
Doctor Scientiae

VIÇOSA - MINAS GERAIS
2025

FELIPE DE OLIVEIRA DIAS

Aplicação de técnicas estatísticas e do sensoriamento remoto no melhoramento genético do tomateiro para produtividade, qualidade e resistência à requeima

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

Orientador: Carlos Nick Gomes

Coorientadores: Domingos S. M. Valente
Kaio O. das Gracas Dias

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
D541a
2025

Dias, Felipe de Oliveira, 1995-
Aplicação de técnicas estatísticas e do sensoriamento remoto no melhoramento genético do tomateiro para produtividade, qualidade e resistência à requeima / Felipe de Oliveira Dias. – Viçosa, MG, 2025.
1 tese eletrônica (88 f.): il. (algumas color.).

Orientador: Carlos Nick Gomes.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2025.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2025.478
Modo de acesso: World Wide Web.

1. Tomate - Melhoramento genético - Métodos estatísticos. 2. Tomate - Melhoramento genético - Sensoriamento remoto. 3. Aprendizado do computador. I. Gomes, Carlos Nick, 1979-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Fitotecnia. III. Título.

CDD 22. ed. 635.6422

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

FELIPE DE OLIVEIRA DIAS

Aplicação de técnicas estatísticas e do sensoriamento remoto no melhoramento genético do tomateiro para produtividade, qualidade e resistência à requeima

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

APROVADA: 12 de junho de 2025.

Assentimento:

Felipe de Oliveira Dias
Autor

Carlos Nick Gomes
Orientador

Essa tese foi assinada digitalmente pelo autor em 24/07/2025 às 14:39:55 e pelo orientador em 25/07/2025 às 07:35:56. As assinaturas têm validade legal, conforme o disposto na Medida Provisória 2.200-2/2001 e na Resolução nº 37/2012 do CONARQ. Para conferir a autenticidade, acesse https://siadoc.ufv.br/validar-documento. No campo 'Código de registro', informe o código **17FK.RK9S.93UF** e clique no botão 'Validar documento'.

Aos meus pais Osvando (*in memoriam*) e Terezinha;
Aos meus irmãos Silvano, Eliana, Elizangela, Geovani, Rosa, Renata, Fabiana, Tatiane e Bruna;
Aos meus afilhados Alice, Tiago e Sofia;
Aos meus sobrinhos Lucas, Gustavo, Fernanda, Júlia, Miguel, José e Helena.

Dedico

AGRADECIMENTOS

Agradeço a Deus,
Agradeço aos meus familiares,
Agradeço aos meus amigos,
Agradeço aos técnicos de campo, laboratório e administrativos,
Agradeço aos bolsistas e estagiários,
Agradeço aos colegas da pós-graduação,
Agradeço aos professores,
Agradeço aos meus coorientadores,
Agradeço ao meu orientador,
Agradeço ao povo brasileiro,
Agradeço às políticas públicas,
Agradeço ao NEPFit,
Agradeço ao Departamento de Agronomia,
Agradeço ao PPG em Fitotecnia,
Agradeço à UFV,
Agradeço à CAPES.
A todos, muito obrigado!

VIVA A EDUCAÇÃO! VIVA A CIÊNCIA!

Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

“*É meu filho, estudar não deve ser fácil mesmo não. Mas se você falar comigo que quer desistir, depois de tudo o que já gastei contigo, eu vou te dar uma coça que você não vai querer outra.*”
(Terezinha Cezaria de Oliveira Dias)

“*Deus abençoa os passos do Felipe, que ele vai em frente!*”
(Osvando Dias)

RESUMO

DIAS, Felipe de Oliveira, D.Sc., Universidade Federal de Viçosa, junho de 2025. Aplicação de técnicas estatísticas e do sensoriamento remoto no melhoramento genético do tomateiro para produtividade, qualidade e resistência à requeima . Orientador: Carlos Nick Gomes. Coorientadores: Domingos Sarvio Magalhaes Valente e Kaio Olimpio das Gracas Dias.

A produção vegetal sustenta o estilo de vida das sociedades fornecendo alimentos, fibras e energia. A agricultura deve atender à crescente demanda por esses produtos, superando os desafios do clima que se impõe. O melhoramento genético vegetal é uma ciência motora da evolução da agricultura, desenvolvendo cultivares mais produtivas e adaptadas a diferentes ambientes. Muitas técnicas podem ser utilizadas na avaliação e seleção de genótipos superiores, visando alcançar objetivos diversos. O tomate é um alimento muito apreciado no mundo, e deve ser melhorado geneticamente para obter melhor performance agronômica. Um dos objetivos do melhoramento do tomateiro é obter cultivares produtivas e com qualidade do fruto. Assim, o primeiro capítulo desse trabalho visou selecionar genótipos superiores com maior produtividade e qualidade do fruto utilizando a técnica de seleção para múltiplas características através do índice de seleção FAI-BLUP. Para melhorar a acurácia seletiva, utilizou-se a técnica de modelagem da dependência espacial de experimentos de campo para selecionar o modelo de melhor ajuste para predizer os valores genotípicos para cada característica. Considerou-se três modelos de ajuste: o modelo 1 foi o modelo tradicional de análise de experimentos no delineamento de blocos casualizados, que presume independência espacial para os efeitos residuais; o modelo 2 considerou as correlações autorregressivas de primeira ordem (AR1) nas direções de linha e coluna conforme o posicionamento das parcelas experimentais no campo; e o modelo 3 que além de considerar as correlações AR1 em ambas as direções, considerou também a variação adicional independente entre parcelas (efeito nugget). Procedeu-se a seleção do modelo de melhor ajuste para cada característica pelos critérios de AIC, BIC e acurácia. Implementou-se o índice FAI-BLUP em dois cenários: modelando a dependência espacial (Cenário I) e sem a modelagem dessa dependência (Cenário II). O Cenário I foi o mais adequado para se obter ganhos genéticos conforme esperado, em direção ao ideótipo delineado. Além de obter cultivares produtivas e com qualidade do fruto, estas devem ser resistentes as principais pragas e doenças do tomateiro. Assim, o segundo capítulo desse trabalho buscou

utilizar técnicas de sensoriamento remoto para avaliar genótipos de tomate em campo quanto a severidade da requeima, uma das principais doenças dessa cultura. Imagens multiespectrais do campo experimental foram obtidas com a câmera MicaSense RedEdge-MX a bordo do drone DJI, Matrice 100. As imagens foram processadas para extrair os dados das bandas espectrais e calcular os índices de vegetação. Estes índices foram utilizados para treinar modelos preditivos de aprendizado de máquinas de modo supervisionado, utilizando o algoritmo Random Forest. Duas metodologias foram consideradas: o Método 1 utilizou os índices de vegetação calculados a partir de uma única imagem obtida no último dia de avaliação; o Método 2 utilizou os índices de vegetação calculados a partir de quatro imagens obtidas em distintos dias de avaliação. Os modelos treinados com o Método 2 tiveram melhor desempenho preditivo. Os genótipos avaliados tiveram a severidade da requeima predita pelo melhor modelo selecionado após seu treinamento. As testemunhas resistentes e suscetível foram adequadamente classificadas pela severidade predita, possibilitando a seleção de genótipos de tomateiro resistentes a requeima.

Palavras-chave: *Solanum lycopersicum*; acurácia seletiva; análise espacial; índice FAI-BLUP; imagem multiespectral; aprendizado de máquinas

ABSTRACT

DIAS, Felipe de Oliveira, D.Sc., Universidade Federal de Viçosa, June, 2025.
Application of statistical techniques and remote sensing in the genetic improvement of tomato crop for productivity, quality and resistance to late blight
. Adviser: Carlos Nick Gomes. Co-advisers: Domingos Sarvio Magalhaes Valente and Kaio Olimpio das Gracas Dias.

Plant production sustains society's lifestyle by providing food, fiber, and energy. Agriculture must meet the growing demand for these products while overcoming the challenges of climate change. Plant breeding is a science driving the evolution of agriculture, developing more productive cultivars adapted to different environments. Many techniques can be used to evaluate and select superior genotypes to achieve various objectives. Tomatoes are a highly valued food worldwide and must be genetically improved to achieve better agronomic performance. One of the objectives of tomato breeding is to obtain productive cultivars with high fruit quality. Therefore, the first chapter of this work aimed to select superior genotypes with higher productivity and fruit quality using the multi-trait selection technique using the FAI-BLUP selection index. To improve selective accuracy, we used the spatial dependence modeling technique for field experiments to select the best-fitting model for predicting genotypic values for each trait. Three adjustment models were considered: Model 1 was the traditional model for analyzing experiments in a randomized block design, which assumes spatial independence for residual effects; Model 2 considered first-order autoregressive correlations (AR1) in the row and column directions according to the positioning of the experimental plots in the field; and Model 3, which, in addition to considering AR1 correlations in both directions, also considered the additional independent variation between plots (the nugget effect). The best-fitting model for each trait was selected using AIC, BIC, and accuracy criteria. The FAI-BLUP index was implemented in two scenarios: modeling spatial dependence (Scenario I) and without modeling this dependence (Scenario II). Scenario I was the most appropriate for obtaining genetic gains as expected, toward the delineated ideotype. In addition to obtaining productive cultivars with fruit quality, these cultivars must be resistant to the main tomato pests and diseases. Thus, the second chapter of this work sought to use remote sensing techniques to assess tomato genotypes in the field for the severity of late blight, one of the main diseases of this crop. Multispectral images of the experimental field were obtained with the MicaSense RedEdge-MX camera onboard the DJI Matrice 100 drone. The images were processed to extract spectral band data and calculate vegetation indices. These indices

were used to train supervised machine learning predictive models using the Random Forest algorithm. Two methodologies were considered: Method 1 used vegetation indices calculated from a single image acquired on the last day of evaluation; Method 2 used vegetation indices calculated from four images acquired on different evaluation days. The models trained with Method 2 had better predictive performance. The genotypes evaluated had their late blight severity predicted by the best model selected after training. The resistant and susceptible controls were adequately classified by predicted severity, enabling the selection of tomato genotypes resistant to late blight.

Keywords: *Solanum lycopersicum*; selective accuracy; spatial analysis; FAI-BLUP index; multispectral imaging; machine learning

SUMÁRIO

1. INTRODUÇÃO GERAL

A produção vegetal deve ser dobrada para sustentar a população crescente no mundo, e deve lidar com as constantes mudanças no clima (AZIZ; MASMOUDI, 2024; RESENDE; BRONDANI; CHAVES, 2023). O melhoramento genético de plantas colabora significativamente com esse desafio, sempre inovando com cultivares que visam atender as diferentes necessidades de mercado (ACQUAAH, 2016).

Muitas técnicas são adotadas nos programas de melhoramento vegetal, que englobam todos os processos a fim de melhorar os ganhos genéticos das características de uma cultura, para que objetivos, como a maior produtividade, sejam alcançados (ANAND; SUBRAMANIAN; KAR, 2023). A técnica para ser utilizada em determinado programa de melhoramento depende de muitos fatores como a cultura, o modo de reprodução, a herança das características, o tipo de cultivar, os objetivos, os recursos disponíveis (ACQUAAH, 2016). Todo o processo avaliativo e seletivo das populações de melhoramento pode ser amparado por diferentes técnicas a fim de atingir os objetivos do melhoramento de modo efetivo (ANAND; SUBRAMANIAN; KAR, 2023).

O tomate é um alimento de grande importância no mundo, com produção total de quase 200 milhões de toneladas em 2023 (FAO, 2025). A versatilidade de uso na culinária o torna um alimento atrativo em diferentes combinações de pratos. Atrai mais ainda o fato de ser um alimento de baixa caloria e fonte de alguns minerais e vitaminas, colaborando com a tão desejada dieta saudável e nutritiva em benefício da saúde (DORAIS; EHRET; PAPADOPOULOS, 2008).

Os programas de melhoramento genético do tomateiro possuem objetivos recorrentes para lidar com os desafios enfrentados na tomaticultura. Assim, deve-se ter em mente o desenvolvimento de cultivares com alta produtividade e qualidade do fruto, com resistência as principais pragas e doenças e adaptadas ao ambiente de cultivo (BERGOUGNOUX, 2014).

Em programas de melhoramento genético vegetal, a seleção de genótipos superiores considerando múltiplas características de interesse agronômico é realizada por meio de índices de seleção. No entanto, problemas de colinearidade entre características nos dados podem comprometer a eficiência dos índices lineares tradicionais (NIRMALARUBAN et al., 2025). O índice de seleção FAI-BLUP supera essa limitação ao utilizar análise exploratória de fatores, que produz eixos ortogonais

não correlacionados entre os fatores finais, reduzindo a dimensionalidade dos dados. Assim, características agrupadas no mesmo fator são correlacionadas enquanto aquelas agrupadas em fatores diferentes são não correlacionadas (ROCHA; MACHADO; CARNEIRO, 2018). O FAI-BLUP mantém dessa forma a natureza das correlações genéticas entre as características, que é essencial em programas de melhoramento modernos (NIRMALARUBAN et al., 2025; ROCHA; MACHADO; CARNEIRO, 2018).

O sensoriamento remoto por meio de drones vem sendo amplamente empregado nos programas de melhoramento genético, possibilitando a fenotipagem de alto rendimento dos genótipos sob seleção (MOCHIDA et al., 2018). As imagens obtidas via sensoriamento remoto podem reduzir a subjetividade das avaliações fenotípicas devido aos erros humanos, aumentando a acurácia da seleção (SINGH et al., 2021). Dessa forma, abre-se caminho para implementar metodologias mais eficientes nos programas de melhoramento, visando o desenvolvimento de cultivares mais produtivas e mais adaptadas aos sistemas agrícolas (RESENDE; BRONDANI; CHAVES, 2023).

Considerando essas técnicas que podem ser empregadas no melhoramento genético do tomateiro, este trabalhou objetivou utilizar o índice de seleção FAI-BLUP (ROCHA; MACHADO; CARNEIRO, 2018) associado a modelagem da dependência espacial para selecionar genótipos superiores para a produção e qualidade do fruto (Capítulo 1); e sensoriamento remoto associado ao uso do aprendizado de máquinas para predizer a severidade da requeima para auxiliar a seleção de genótipos resistentes (Capítulo 2). Ambos os capítulos são precedidos de uma revisão de literatura.

2. REVISÃO DE LITERATURA

2.1. O tomateiro

O tomateiro cultivado é conhecido cientificamente como *Solanum lycopersicum* L. var. *lycopersicum* (sinônimo *Lycopersicon esculentum* Mill.). Popularmente o termo tomate é adotado amplamente para se referir a cultura, nome dado aos seus frutos. O gênero *Solanum* abrange o maior número de espécies da família Solanaceae, e possui táxons em todos os continentes com alta diversidade, com espécies utilizadas

para diversos usos e, portanto, o gênero de maior importância econômica nessa família botânica (BERGOUGNOUX, 2014).

A história da domesticação do tomate é descrita do seguinte modo: aproximadamente 78000 anos atrás ocorreu na região do Equador a especiação natural com *S. pimpinellifolium* dando origem a *S. lycopersicum* var. *cerasiforme* sem intervenção humana; nessa região, cerca de 3000 a 2000 anos a.C, iniciou-se a seleção de plantas de *S. lycopersicum* var. *cerasiforme* pelos Povos Maias para características de frutos; biótipos já selecionados chegaram a região do México, possivelmente como erva daninha, e tiveram sua domesticação completada pelos Povos Astecas em *S. lycopersicum* var. *lycopersicum*, cerca de 700 anos d.C.; o contato europeu com estes povos na época das Grandes Navegações possibilitou o então tomate domesticado chegar a Europa, disseminando-o para o restante do mundo (BLANCA et al., 2012, 2015; FLORES et al., 2024). Durante a domesticação, a seleção ocorreu principalmente para a massa do fruto, enquanto o formato teve importância secundária (BLANCA et al., 2015).

O tomateiro é cultivado como uma planta anual, de porte rasteiro a arbustivo, com crescimento determinado a indeterminado, raiz do tipo pivotante, caule herbáceo e piloso, com folhas compostas e pilosas, inflorescência do tipo cacho, flores hermafroditas, ovário bi ou plurilocular, autógama, fruto do tipo baga, sementes dicotiledôneas e germinação epígea (ALVARENGA, 2013).

O tomate pode ser utilizado em diversos pratos na culinária de modo *in* natura ou utilizado na indústria de processamento. É um alimento com baixo poder calórico, fonte de minerais como cálcio e potássio, de vitaminas A e C, e de carotenoides como o licopeno. O consumo de tomate pode auxiliar na prevenção de doenças cardiovasculares e alguns tipos de câncer (DORAIS; EHRET; PAPADOPOULOS, 2008). Compostos bioativos, com propriedades nutracêuticas são relatados inclusive para suas sementes (KUMAR et al., 2021).

2.2. Panorama mundial da produção de tomate

A Figura 1 apresenta um retrato da área destinada ao cultivo de tomate no mundo nos anos 2000 e 2023. De acordo com dados da Organização das Nações Unidas para a Alimentação e a Agricultura (FAO, 2025), nos anos 2000 uma área de aproximadamente quatro milhões de hectares foi utilizada para este fim, enquanto nos anos 2023 essa área ultrapassa os cinco milhões de hectares. A diferença entres os

dois anos resulta em um incremento de 40% na área cultivada. China, Índia e Nigéria são os países que lideram em valores totais a área utilizada para a tomaticultura no mundo. No Brasil, a área destinada ao cultivo obteve um suave aumento de 56 para 59 mil hectares entre os dois anos.

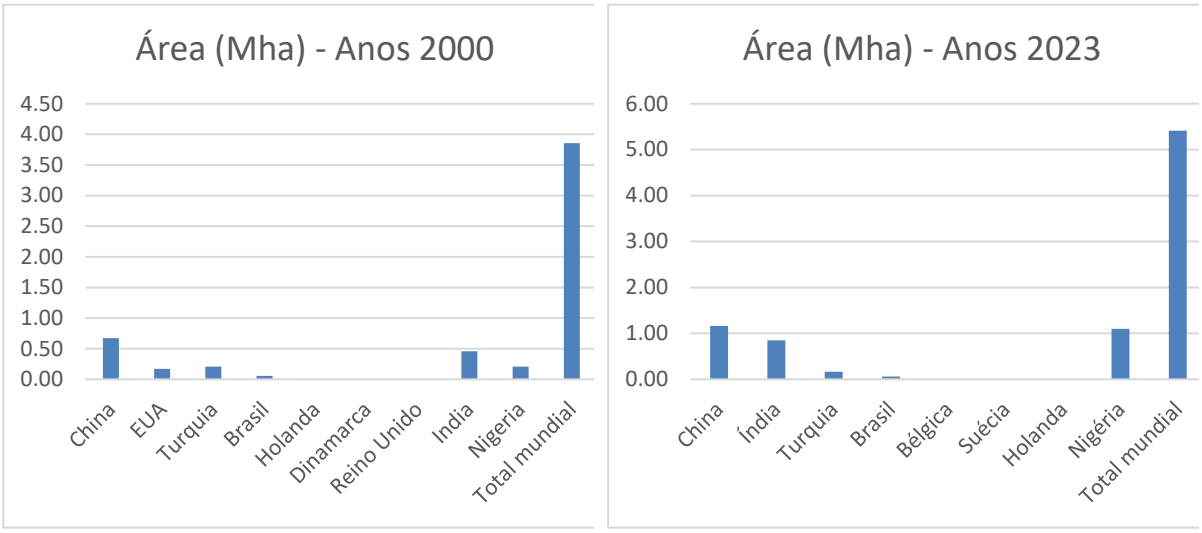

Figura 1: Retrato da área cultivada com tomate no mundo nos anos 2000 e 2023, em milhões de hectares (Mha). Dados: FAO, 2025.

A Figura 2 apresenta um retrato da produção total de tomate no mundo para os anos 2000 e 2023, conforme FAO (2025). Nos anos 2000 a produção global foi de pouco mais de 100 milhões de toneladas (Mt), enquanto nos anos 2023 a produção foi de quase 200 Mt. A diferença entre os dois anos representa um incremento de 75% na produção total em todo o mundo. Nos anos 2000, China, EUA e Turquia lideravam a produção mundial, enquanto nos anos 2023 China, Índia e Turquia se configuraram como os maiores produtores. O Brasil se posicionava como nono maior produtor mundial nos anos 2000, com produção estimada de 3 Mt. Nos anos 2023 o Brasil foi o oitavo maior produtor, produzindo mais de 4 Mt.

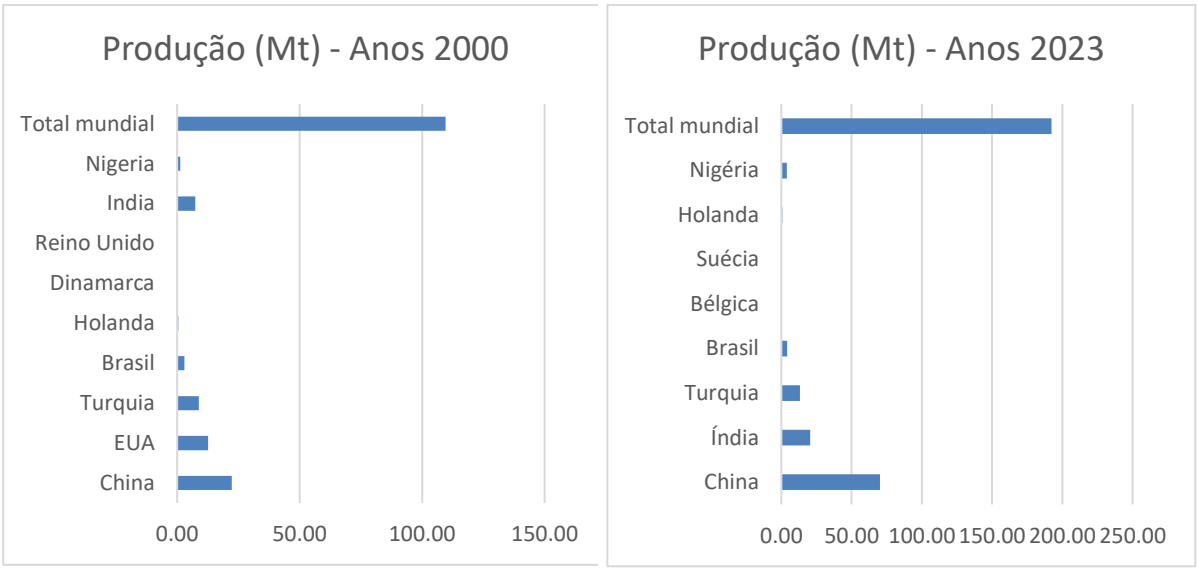

Figura 2: Retrato da produção mundial de tomate no mundo nos anos 2000 e 2023, em milhões de toneladas (Mt). Dados: FAO, 2025.

A Figura 3 apresenta o retrato do rendimento produtivo de tomate no mundo para os anos 2000 e 2023 (FAO, 2025). O rendimento médio mundial da produção de tomate passou de 42 para 60 toneladas por hectare (t/ha) entre os anos 2000 e 2023. Essa diferença representa um incremento de 42%. Holanda, Dinamarca e Reino Unido lideravam a produtividade de tomate nos anos 2000, enquanto nos anos 2023, a liderança fica com Bélgica, Suécia e Holanda. O rendimento brasileiro aumentou de 52 para 70 t/há entre os dois anos, acima da média mundial.

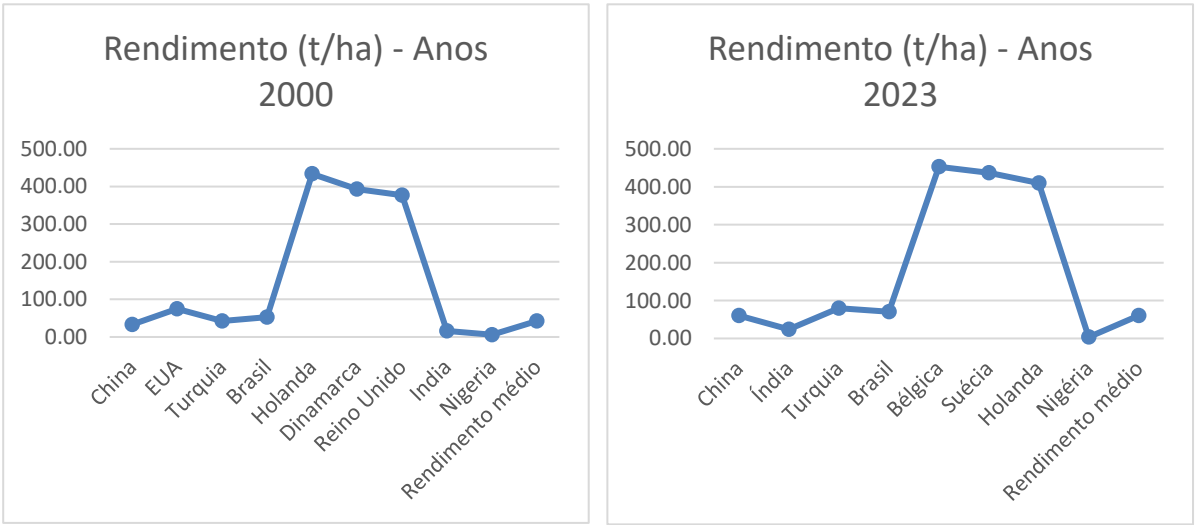

Figura 3: Retrato do rendimento do cultivo de tomate no mundo nos anos 2000 e 2023, em toneladas por hectare (t/ha). Dados: FAO, 2025.

Os países que apresentam o maior rendimento ultrapassam a produtividade de 400 t/ha, demonstrando o potencial produtivo que a tomaticultura pode alcançar. Ao

mesmo tempo, países como a Nigéria que apresenta rendimento muito aquém da média mundial, indica que existe muitos fatores a serem superados para melhorar estas estatísticas no país. Nota-se que os países que possuem maior rendimento não possuem condições climáticas favoráveis a tomaticultura (BERGOUGNOUX, 2014).

A produtividade média de tomate no Brasil ainda é muito baixa se comparado aos líderes globais, fazendo-se necessário melhorar as técnicas de cultivo e obter cultivares com melhor performance agronômica, ou seja, mais produtivas e mais adaptadas as condições de cultivo.

Segundo o IBGE (2024), no Brasil os maiores produtores de tomate no ano de 2024 foram os estados de Goiás, São Paulo e Minas Gerais com participação de 31; 23 e 12% da produção nacional, respectivamente. Em Goiás, a produção é voltada principalmente a indústria de processamento, enquanto em São Paulo e Minas Gerais é mais voltada ao consumo *in* natura. A tomaticultura demanda alto investimento, sendo uma das atividades com maior custo financeiro na olericultura. Ainda assim, houve um incremento na produção total entre os maiores produtores brasileiros entre os anos 2023 e 2024 (IBGE, 2024).

De acordo com dados da Companhia Nacional de Abastecimento (CONAB, 2025), as estimativas do custo de produção de tomate no município de Coimbra – MG em 2023 foi de aproximadamente R$150.000,00/ha. Lideraram os custos as despesas com embalagens e utensílios (R$45.000,00/ha), fertilizantes (R$38.000,00/ha) e mão de obra (R$37.000,00/ha). Os dados foram estimados baseando-se no cultivo convencional irrigado de agricultores familiares, com produtividade média esperada de 100 t/ha. Dessa forma, obter cultivares mais produtivas pode também auxiliar na diluição dos custos de produção e favorecer o consumo de tomate pela redução do preço final ao consumidor.

O impacto social da tomaticultura é altamente relevante, uma vez que pode empregar diretamente entre quatro e cinco trabalhadores por hectare, além de toda cadeia gerada além do campo com empregos diretos e indiretos (TREICHEL, 2016).

Mais de 400 municípios brasileiros são responsáveis pelo abastecimento das CEASAS com tomate em todo país (CONAB, 2024). Os cultivos já estão dominados com as cultivares híbridas, seja para consumo fresco ou destinado a indústria de processamento (CONAB, 2019). Apesar de pouco expressivo, o Brasil também exporta tomates de mesa para Argentina, Uruguai e Paraguai que se beneficia da proximidade entre os países, garantindo que os frutos cheguem em boas condições

de consumo. Corrobora também o fato da ausência de imposto de importação e exportação entre os países membro do Mercosul, que favorece a redução no custo final ao consumidor (CONAB, 2019).

2.3. Melhoramento genético do tomateiro

O tomateiro é uma espécie diploide com 2n = 2X = 24 cromossomos, considerado de base genética estreita quando comparado aos seus ancestrais silvestres (NUEZ; DÍEZ, 2013). Estima-se que menos de 5% da variação genética encontrada nas espécies silvestres está presente em *S. lycopersicum* var. *lycopersicum* (BAI; LINDHOUT, 2007).

Obter cultivares mais adaptadas ao ambiente, com resistência as principais pragas e doenças, com alto rendimento e qualidade do fruto são os objetivos recorrentes no melhoramento genético do tomate (BERGOUGNOUX, 2014). Nesse sentido, as espécies silvestres de tomate podem contribuir significativamente com os objetivos do melhoramento, com muitas características desejáveis já identificadas nessas espécies (KLEE; RESENDE, 2020). As espécies *S. pimpinellifolium*, *S. peruvianum*, *S. habrochaites* e *S. chilense* são consideradas as melhores fontes de resistência para vírus, fungos e bactérias (NUEZ; DÍEZ, 2013). Já para melhorar a qualidade do fruto, as espécies *S. pimpinellifolium* e *S. chmielewskii* se destacam (BERGOUGNOUX, 2014).

Apesar dos esforços dos melhoristas em desenvolver cultivares de tomate adaptadas a inúmeras condições de cultivo no mundo, as mudanças climáticas representam um novo desafio ao melhoramento da cultura (KLEE; RESENDE, 2020). Entre os estresses abióticos, estudos voltados a salinidade, a seca e a temperaturas extremas têm sido mais frequentes (NUEZ; DÍEZ, 2013). Para esse propósito, as espécies *S. cheesmaniae*, *S. pennellii* e *S. habrochaites* são consideradas boas fontes de recursos genéticos (BERGOUGNOUX, 2014).

Mesmo com todas essas fontes potenciais de recursos genéticos, a principal desvantagem da utilização das espécies silvestres nos programas de melhoramento é o arraste gênico de características indesejadas (VU et al., 2020). Nesse caso, faz-se necessário a adoção do pré-melhoramento para introgressão das características de interesse das espécies silvestres em linhagens elite do tomateiro cultivado (ACQUAAH, 2016). Muitas gerações são requeridas para eliminar as características indesejadas e obter linhagens estáveis (AZIZ; MASMOUDI, 2024). Atualmente, com

as técnicas de engenharia genética esse problema pode ser contornado mais rapidamente. A edição gênica é uma esperança para o futuro da agricultura mundial, uma vez que possibilita a expressão e o silenciamento controlados de genes, viabilizando assim a domesticação *de novo* do tomateiro em um menor tempo (VU et al., 2020).

Nessa perspectiva, Zsögön et al. (2018) utilizaram a técnica CRISPR-Cas9 para editar geneticamente *S. pimpinellifolium* com a finalidade de sua domesticação *de novo*. Os autores editaram seis loci e obtiveram plantas editadas do referido tomateiro silvestre com um aumento três vezes maior no tamanho dos frutos e dez vezes maior na quantidade de frutos quando comparadas as plantas não editadas. Em comparação as plantas de *S. lycopersicum*, as plantas editadas de *S. pimpinellifolium* tiveram um aumento de 500% no teor de licopeno. Todas essas manipulações foram feitas em uma única geração com um único experimento de transformação, abrindo caminho para novos avanços no melhoramento molecular do tomateiro.

A evolução do melhoramento genético vegetal possui as seguintes fases, conforme Resende, Brondani e Chaves (2023): o Melhoramento Genético 1.0, marcado pelos eventos de domesticação das plantas até o surgimento da agricultura; o Melhoramento Genético 2.0, marcado pelas Leis de Mendel na área da genética e pela biometria, possibilitando previsão dos ganhos com a seleção e sistematizando os processos de melhoramento; o Melhoramento Genético 3.0, marcado pelas técnicas moleculares e genômicas que possibilitaram a identificação de múltiplos efeitos de genes associados aos fenótipos, acelerando e aumentando o rigor no desenvolvimento de cultivares; e o Melhoramento Genético 4.0, em grande expansão atualmente, com adoção de tecnologias como a inteligência artificial, sistemas de informações geográficas e sensoriamento remoto, proporcionando grandes avanços na genotipagem e fenotipagem de alto rendimento.

Considerando esse avanço na área do melhoramento genético, Burgueño (2018) ressalta o desenvolvimento computacional e de softwares para implementar metodologias estatísticas e matemáticas mais aprimoradas. Com isso, muitas análises podem serem feitas mais rapidamente no mesmo conjunto de dados para selecionar-se a melhor, como é o caso das análises estatísticas espaciais.

2.3.1. Análise estatística espacial de experimentos de campo

Pressupõe-se independência de erros em análises tradicionais de experimentos de campo, ou seja, é esperado que as observações de parcelas experimentais vizinhas são não correlacionadas (RESENDE; STURION, 2001). No entanto, a variabilidade irregular nos diversos componentes do ambiente pode afetar essa pressuposição. No planejamento experimental, essa variabilidade é considerada e alguns delineamentos são utilizados para controlá-la. Mas, esses delineamentos falham em controlar todas as fontes de variação possíveis de estarem presentes no campo experimental, e mesmo dentro de blocos a princípio homogêneos, variabilidade intrabloco pode existir (BURGUEÑO, 2018).

O controle do efeito ambiental do campo experimental é fundamental para uma comparação mais acurada entre genótipos e capitalização efetiva dos efeitos genéticos. Esse controle, além de poder ser realizado por meio dos delineamentos, pode também ser feito via análise levando em conta o posicionamento das parcelas para corrigir as tendências espaciais (HOEFLER et al., 2020). De acordo com Gilmour, Cullis e Verbyla (1997), as tendências espaciais podem ser do tipo local e global, devido a padrões que podem ser, respectivamente, específicos ou irregulares nas características físicas, químicas e biológicas do solo, condições climáticas ou ainda serem introduzidas pelas práticas de manejo da cultura ao longo da experimentação agrícola.

Na estrutura de modelos mistos, as tendências locais são ajustadas via processos autorregressivos separáveis de primeira ordem (AR1), que pode se na direção de linhas, colunas ou ambos; enquanto as tendências globais são ajustadas pelo ajuste de parâmetros extras como polinômios ou splines (VELAZCO et al., 2017). Para isso, as coordenadas espaciais em termos de linha e coluna das parcelas experimentais devem ser conhecidas para o processo de modelagem das tendências (GILMOUR; CULLIS; VERBYLA, 1997).

Em experimentos conduzidos em um único ambiente, enfatiza-se a análise genética em detrimento da análise ambiental, classificando os genótipos por seus valores genéticos para realizar a seleção. Uma vez que os valores genéticos são baseados nos valores fenotípicos e este último pode sofrer influência ambiental, o efeito do ambiente deveria ser melhor estudado, sendo a abordagem estatística de análise espacial uma importante ferramenta nesse contexto (RESENDE; STURION,

2001). Assim, a análise espacial é um método de controle local posterior a implementação do experimento, usada para melhorar a acurácia das predições e a comparação de genótipos (BURGUEÑO, 2018).

2.3.2. Índice de seleção FAI-BLUP

No melhoramento genético vegetal, a seleção de genótipos para múltiplas características de interesse agronômico é feita utilizando índices de seleção. Os índices de seleção tradicionais, como Smith-Hazel, podem falhar ao se considerar muitas características e suas relações específicas devido a problemas de multicolinearidade (NIRMALARUBAN et al., 2025). Assim, havendo multicolinearidade na matriz de covariância fenotípica entre as características, os índices tradicionais não devem ser usados, uma vez que os parâmetros podem não serem corretamente estimados, os erros podem aumentar nessas estimativas e enviesar as estatísticas de inferência com conclusões não confiáveis (ROCHA et al., 2019).

Para resolver o problema da multicolinearidade, estuda-se a correlação entre características e elimina-se uma de duas características altamente correlacionadas entre si. Assim, índices de seleção tradicionais consideram poucas características para simplificar o problema estatístico, mas informações importantes podem ser desconsideradas (ROCHA; MACHADO; CARNEIRO, 2018). Para o índice Smith-Hazel, outro problema enfrentado pelos melhoristas é a atribuição de pesos para cada característica conforme sua importância, e tenta-se traduzir essa ponderação baseado em valores econômicos (OLIVOTO; NARDINO, 2021).

O índice FAI-BLUP foi proposto para superar esses problemas e incorporar todas as características disponíveis na seleção. Nesse índice, as correlações genéticas entre as características e o ideótipo são consideradas (ROCHA et al., 2019). O ideótipo pode ser definido como o genótipo que reúne em si todas as características agronômicas desejáveis para uma dada cultura. Assim, os melhoristas se guiam ao longo dos ciclos seletivos no programa de melhoramento, com uma visão de futuro para aumentar o desempenho produtivo e qualidade das cultivares a serem geradas (OLIVOTO; NARDINO, 2021).

No índice FAI-BLUP, a análise de componentes principais é utilizada para extrair as cargas fatoriais da matriz de correlação genética. O número de fatores é então utilizado no cálculo do número de ideótipos, que é uma combinação de fatores desejáveis e indesejáveis. Posteriormente, a distância entre genótipo e ideótipo é

estimada e convertida em probabilidade espacial para ranquear todos os genótipos. Por considerar a estrutura de correlação presente nos dados, o índice FAI-BLUP levará a seleção daqueles genótipos mais próximos do ideótipo delineado pelo melhorista (ROCHA; MACHADO; CARNEIRO, 2018).

2.3.3. Sensoriamento remoto

A coleta de imagens com uso do sensoriamento remoto vem auxiliando a fenotipagem de alto rendimento das culturas agrícolas, de modo não destrutivo (MOCHIDA et al., 2018). O sensoriamento remoto pode ser definido como a técnica de obter informações de um objeto sem interação direta com ele, coletando a radiação eletromagnética refletida e emitida por esse objeto que pode ser armazenada na forma de imagem (HORNING, 2019). Adicionalmente ao registro e monitoramento da energia eletromagnética captada, inclui-se ao conceito de sensoriamento remoto todo o processamento da informação para torná-la útil e aplicável em diferentes áreas do conhecimento (KUMAR; SINGH; KAUR, 2019).

A energia eletromagnética dividida em comprimentos de onda é chamada de espectro eletromagnético, variando de comprimentos de onda curtos (raios gama, com alta frequência) até comprimentos de onda longos (ondas de rádio, com baixa frequência). Outra subdivisão da região espectral são as bandas espectrais, que compartilham características semelhantes entre comprimentos de onda: faixa do visível entre 400 e 700 nm (Blue, Green, Red) e infravermelho variando entre 700 e 10^6 nm (KUMAR; SINGH; KAUR, 2019).

O sensoriamento remoto pode ser do tipo passivo ou ativo: no primeiro caso a principal fonte de energia eletromagnética é o sol, enquanto no segundo o sensor emite sua própria energia. Ao atingir uma superfície, a energia pode ser refletida e detectada por câmeras ou sensores eletrônicos. Outra parte da energia pode ser absorvida e emitida como energia térmica, detectada através de sensores termais (DEFRIES, 2013).

Os objetos alvos de estudo interagem fisicamente de modo diferente com os comprimentos de onda, o que permite a diferenciação de objetos a partir de dados espectrais. A refletância espectral, definida como a razão entre a intensidade de luz refletida e a intensidade de luz incidente em uma superfície, pode diferir para os diversos objetos. O comportamento espectral das plantas, por exemplo, difere do comportamento espectral da areia branca: enquanto a vegetação absorve a banda do

vermelho e reflete a maior parte do infravermelho, a areia reflete a maior parte de ambas as bandas (HORNING, 2019).

Grande parte dos sistemas de sensoriamento remoto tem como produto imagens digitais, composta por uma matriz de elementos de imagens, os pixels, que equivalem a uma determinada área do objeto avaliado. A refletância medida nessa área pelos sensores é traduzida em um número digital em cada pixel (DEFRIES, 2013). Imagens hiperespectrais são obtidas a partir de sensores que captam uma ampla gama das bandas do espectro eletromagnético, enquanto imagens multiespectrais são obtidas por sensores que gravam informações de apenas algumas bandas (HORNING, 2019).

A resolução espacial e a resolução radiométrica são características importantes das imagens que têm consequências importantes na quantidade de informação disponível. A resolução espacial diz respeito ao tamanho do pixel projetado na superfície, ou seja, a área real que cada pixel representa. Por exemplo, uma resolução de 15 m significa que cada pixel representa uma área real de 15 x 15 m. A resolução radiométrica refere-se aos números digitais (níveis de brilho) que podem ser usados para representar o pixel, em que quanto maior a intensidade da radiação, maior será o número digital. Assim, quanto maior a faixa dos números digitais do sensor, maior será a possibilidade de discriminar diferenças sutis na imagem (DEFRIES, 2013; HORNING, 2019).

As imagens obtidas pelos sensores estão em seu formato bruto e necessitam de processamento para corrigir distorções decorrentes da aquisição e torná-las mais fiéis ao ambiente captado. Segundo Kumar, Singh e Kaur (2019) o processamento de imagens envolve as seguintes fases descritas a seguir: a correção geométrica, realizada para que as coordenadas espaciais da imagem correspondam as coordenadas espaciais da superfície avaliada, reposicionando os pixels de sua localização inicial para uma grade de referência específica georreferenciada; a correção radiométrica, processo pelo qual os erros e ruídos da imagem são minimizados, devido ao fato do pixel adquirir valores extremos em relação aos pixels vizinhos e gerar pontos muito escuros ou brilhantes, causados pelo espalhamento desigual da radiação na atmosfera que depende do comprimento de onda, da presença de nuvens e inclinação solar; o aprimoramento da imagem, com aplicação de técnicas com a finalidade de aumentar as distinções entre as características observadas, melhorando o contraste entre elas e aumentando a qualidade visual de

imagens de baixa visibilidade; a classificação de imagens para extrair as informações úteis, seja de modo visual baseado na experiência do pesquisador e na qualidade da imagem disponível, ou seja de modo computacional baseado em algoritmos que reconhecem padrões nem sempre observáveis a olho nu.

Em relação as plataformas de sensoriamento remoto, os veículos aéreos não tripulados (VANTs), popularmente chamados de drones, vêm sendo massivamente empregados com diversos fins nos campos agrícolas. Em comparação aos satélites, os drones apresentam as vantagens de menor altitude de coleta de dados, menor custo operacional, avaliações mais frequentes, menor dependência das condições climáticas e maior resolução espacial (SHI et al., 2016). Ressalta-se ainda que a utilização do sensoriamento remoto nas pesquisas agrícolas ou no monitoramento de lavouras deve envolver uma abordagem multidisciplinar de diferentes áreas do conhecimento, possibilitando novas e melhores soluções aos desafios do campo (KUMAR; SINGH; KAUR, 2019; SHI et al., 2016).

No melhoramento genético vegetal, o sensoriamento remoto pode ser utilizado no estudo de diversidade genética, na caracterização fenotípica ao longo dos ciclos de melhoramento, na avaliação da adaptação de cultivares ao ambiente e obtenção de informações genéticas (ARAUS; CAIRNS, 2014). Tal técnica possibilita a fenotipagem de alto rendimento, em que centenas a milhares de plantas são avaliadas em um único dia de forma mais precisa e acurada (MIR et al., 2019). Além disso, o sensoriamento, remoto e proximal, pode ser uma alternativa para avaliação de características fenotípicas complexas, difíceis de mensurar ou que exigem amostragem destrutiva (PARMLEY et al., 2019). Um exemplo clássico são aquelas relacionadas às raízes de plantas, como a profundidade, o comprimento e a biomassa. Nesse caso, sensores proximais ativos são mais adequados para este tipo de avaliação, como a tomografia de resistividade elétrica e o radar de penetração no solo (ARAUS; CAIRNS, 2014).

2.3.4. Aprendizado de máquinas

A tecnologia empregada nas pesquisas agrícolas possibilitou a aquisição de grandes conjuntos de dados a respeito dos fenótipos e genótipos de plantas. Com a complexidade desses dados, esforços foram empenhados para extrair significado biológico e explorar seu potencial uso. Nesse contexto, o aprendizado de máquinas

evoluiu com sua ampla utilização na fenotipagem e genotipagem de plantas, sendo tecnicamente usado na análise desses conjuntos de dados (VAN DIJK et al., 2021).

As metodologias tradicionais de análise de dados, como a análise de regressão, podem ser limitadas à medida que a complexidade e dimensionalidade dos dados aumenta. Com o aprendizado de máquinas, as metodologias de seleção de características minimizam o número de variáveis preditoras da variável resposta sem grandes prejuízos para o desempenho dos modelos (PARMLEY et al., 2019).

No campo da inteligência artificial, o aprendizado de máquinas é uma abordagem computacional, com base em métodos estatísticos e probabilísticos, que aprende com determinado conjunto de dados a respeito de um fenômeno específico, podendo classificá-lo, reconhecer padrões e prever tendências futuras (SHAKOOR; LEE; MOCKLER, 2017).

Os algoritmos de aprendizado de máquinas podem executar tarefas de modo supervisionado e não supervisionado. No primeiro caso, os modelos preditivos são treinados com dados conhecidos (dados de treino) que são rotulados com uma classificação conhecida. Obtidos os modelos treinados, estes são usados para fazer previsões em dados não conhecidos (dados de teste). Já o aprendizado não supervisionado busca reconhecer padrões em um conjunto de dados fornecido sem sua rotulagem inicial, agrupando os dados em classes discretas (VAN DIJK et al., 2021).

A modelagem em aprendizado de máquinas pode ser do tipo discriminativa ou generativa. No primeiro caso, a modelagem é utilizada para a classificação de dados, aprendendo a relação entre entrada e saída para prever a classe de novos dados. Já na modelagem generativa, o algoritmo busca aprender como os dados foram gerados, podendo gerar novos dados semelhantes aos de entrada. Para a fenotipagem de plantas, a modelagem discriminativa é mais utilizada (SHAKOOR; LEE; MOCKLER, 2017).

O Random Forest é um algoritmo de aprendizado de máquinas que utiliza o Método Ensemble, em que vários modelos são combinados na construção de um modelo preditor final (VAN DIJK et al., 2021). Neste caso, várias árvores de decisão são criadas, e o treinamento de cada árvore ocorre por um subconjunto de dados selecionados aleatoriamente, tornando nula a correlação entre árvores. Desta forma, o modelo final treinado é baseado em várias regras estabelecidas em cada árvore de decisão, aumentando a robustez da predição. O algoritmo pode ser utilizado na

classificação de dados de modo supervisionado, sendo considerado um dos melhores métodos preditivos em vários estudos (PARMLEY et al., 2019).

Para um bom desempenho dos modelos de aprendizado de máquinas, os dados devem ser apropriadamente coletados com mínimos erros de mensuração. No caso do aprendizado supervisionado, os dados ainda devem ser corretamente rotulados. Os dados também podem ser processados previamente para uma melhor performance de utilização pelos algoritmos (VAN DIJK et al., 2021).

O aprendizado de máquinas possibilita a automatização da detecção de doenças de plantas, podendo ser utilizado diversos algoritmos que vêm sendo testados em diversos estudos. Quanto maior e mais complexo o conjunto de dados, mais viável se torna a sua utilização, com análises significativas para a pesquisa agrícola (SHAKOOR; LEE; MOCKLER, 2017). Ademais, associado a outras técnicas, o uso do aprendizado de máquinas pode acelerar o processo de obtenção de cultivares, inclusive com cultivares mais resilientes ao clima (FAROOQ et al., 2024).

3. REFERÊNCIAS

AZIZ, Mughair Abdul; MASMOUDI, Khaled. Molecular Breakthroughs in Modern Plant Breeding Techniques. **Horticultural Plant Journal**, v. 11, n. 1, p. 15–41, jan. 2024.

ACQUAAH, G. Conventional plant breeding principles and techniques. Em: AL-KHAYRI, J.; JAIN, S. M.; JOHNSON, D. V. (Eds.) **Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools.** [s.l.] Springer International Publishing, 2016. v. 1p. 115–158.

ALVARENGA, M. A. R. Origem, botânica e descrição da planta. Em: ALVARENGA, M. A. R. **Tomate: Produção em campo, casa de vegetação e hidroponia**. 2. ed. Lavras: Editora Universitária de Lavras, 2013. p. 11-21.

ANAND, Achala; SUBRAMANIAN, Madhumitha; KAR, Debasish. Breeding techniques to dispense higher genetic gains. **Frontiers in Plant Science**, v. 13, n. 1076094, 19 jan. 2023.

ARAUS, J. L.; CAIRNS, J. E. Field high-throughput phenotyping: The new crop breeding frontier. **Trends in Plant Science**, v. 19, n. 1, p. 52–61, jan. 2014.

BAI, Y.; LINDHOUT, P. Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? **Annals of Botany**, v. 100, n. 5, p. 1085–1094, out. 2007.

BERGOUGNOUX, V. The history of tomato: From domestication to biopharming. **Biotechnology Advances**, v. 32, n. 1, p. 170–189, jan. 2014.

BLANCA, J. et al. Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. **PLoS ONE**, v. 7, n. 10, 31 out. 2012.

BLANCA, J. et al. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. **BMC Genomics**, v. 16, n. 1, 12 dez. 2015.

BURGUEÑO, J. Spatial Analysis of Field Experiments. Em: GLAZ, B.; YEATER, K. M. (Eds.). **Applied Statistics in Agricultural, Biological, and Environmental Sciences**. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, 2018. p. 319–344.

CONAB – Companhia Nacional de Abastecimento. **Compêndio de Estudos da Conab**. 2019. Disponível em: https://www.conab.gov.br/institucional/publicacoes/compendio-de-estudos-da-conab/item/download/29586_4fe6dd2c9c6d1fa5e1cbc5f82061717d. Acesso em: 01 fev. 2025.

CONAB – Companhia Nacional de Abastecimento. **Boletim Hortigranjeiro**. 2024. Disponível em: https://www.conab.gov.br/info-agro/hortigranjeiros-prohort/boletim-hortigranjeiro. Acesso em: 01 fev. 2025.

CONAB – Companhia Nacional de Abastecimento. Planilhas de custos de produção. Brasília, DF, [s.d.]. Disponível em: https://www.gov.br/conab/pt-br/atuacao/informacoes-agropecuarias/custos-de-producao/planilhas-de-custos-de-producao. Acesso em: 01 fev. 2025.

DEFRIES, Ruth. Remote Sensing and Image Processing. *In*: LEVIN, Simon A. (Org.). **Encyclopedia of Biodiversity: Second Edition**. *[S.l.]*: Academic Press, 2013. v. 6 p. 389–399.

DORAIS, Martine; EHRET, David L.; PAPADOPOULOS, Athanasios P. Tomato (Solanum lycopersicum) health components: From the seed to the consumer. **Phytochemistry Reviews**, n. 7, p. 231–250, jul. 2008.

FAROOQ, M. A. et al. Artificial intelligence in plant breeding. **Trends in Genetics**, v. 40, n. 10, p. 891–908, 1 out. 2024.

FLORES, Stalin Sarango et al. The Tomato's Tale: Exploring Taxonomy, Biogeography, Domestication, and Microbiome for Enhanced Resilience. **Phytobiomes Journal**, v. 8, p. 5–20, 2024.

GILMOUR, A. R.; CULLIS, B. R.; VERBYLA, A. P. Accounting for natural and extraneous variation in the analysis of field experiments. **Source: Journal of Agricultural, Biological, and Environmental Statistics**, v. 2, n. 3, p. 269–293, 1997.

HOEFLER, R. et al. Do Spatial Designs Outperform Classic Experimental Designs? **Journal of Agricultural, Biological, and Environmental Statistics**, v. 25, n. 4, p. 523–552, 1 dez. 2020.

HORNING, Ned. Remote Sensing. *In*: FATH, Brian (Org.). **Encyclopedia of Ecology**. Second Edition ed. *[S.l.]*: Elsevier, 2019. v. 4 p. 404–413.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estatísticas da Produção Agrícola: dezembro de 2024**. IBGE, 2024. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/2415/epag_2024_dez.pdf>. Acesso em: 01 fev. 2025.

KLEE, Harry J.; RESENDE, Marcio F. R. Plant Domestication: Reconstructing the Route to Modern Tomatoes. **Current Biology**, v. 30, n. 8, p. R359–R361, 20 abr. 2020.

KUMAR, Dilip; SINGH, R. B.; KAUR, Ranjeet. Remote-Sensing Technology. *In*: KUMAR, Dilip; SINGH, R. B.; KAUR, Ranjeet (Orgs.). **Spatial Information Technology for Sustainable Development Goals**. 1. ed. Cham: Springer, 2019. p. 27–58.

KUMAR, Manoj et al. Tomato (Solanum lycopersicum L.) seed: A review on bioactives and biomedical activities. **Biomedicine and Pharmacotherapy**, n. 142, 1 out. 2021.

MIR, Reyazul Rouf et al. High-throughput phenotyping for crop improvement in the genomics era. **Plant Science**, n. 282, p. 60–72, 1 maio 2019.

MOCHIDA, Keiichi et al. Computer vision-based phenotyping for improvement of plant productivity: A machine learning perspective. **GigaScience**, v. 8, n. 1, p. 1–12, 6 dez. 2018.

NIRMALARUBAN, Rajamani et al. Rooting for resilience: central metaxylem area as a breeding target for yield gain and resilience in wheat (Triticum aestivum L.). **BMC Plant Biology**, v. 25, n. 1, 1 dez. 2025.

NUEZ, F.; DÍEZ, M. J. Solanum lycopersicum var. lycopersicum (Tomato). *In*: MALOY, Stanley; HUGHES, Kelly (Orgs.). **Brenner's Encyclopedia of Genetics**. Second Edition ed. *[S.l.]*: Academic Press, 2013. p. 476–480.

OLIVOTO, T.; NARDINO, M. MGIDI: Toward an effective multivariate selection in biological experiments. **Bioinformatics**, v. 37, n. 10, p. 1383–1389, 15 maio 2021.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A ALIMENTAÇÃO E A AGRICULTURA. FAOSTAT - **Dados Estatísticos**. FAO, 2025. Disponível em: <https://www.fao.org/faostat/en/#data>. Acesso em: 01 fev. 2025.

PARMLEY, K. A. et al. Machine Learning Approach for Prescriptive Plant Breeding. **Scientific Reports**, v. 9, n. 1, 1 dez. 2019.

RESENDE, Marcos Deon Vilela; STURION, José Alfredo. **Análise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de séries temporais empregando REML/BLUP ao nível individual**. 1ª ed. Colombo, PR: Embrapa Florestas, 2001.

RESENDE, Rafael Tassinari; BRONDANI, Claudio; CHAVES, Lazaro José. O melhoramento na era de agricultura de precisão. *In*: RESENDE, Rafael Tassinari; BRONDANI, Claudio (Orgs.). **Melhoramento de Precisão Aplicações e perspectivas na genética de plantas**. 1ª edição ed. Brasília: Embrapa, 2023. p. 13–39.

ROCHA, J. R. DO A. S. DE C. et al. Selection of superior inbred progenies toward the common bean ideotype. **Agronomy Journal**, v. 111, n. 3, p. 1181–1189, 1 maio 2019.

ROCHA, João Romero do Amaral Santos de Carvalho; MACHADO, Juarez Campolina; CARNEIRO, Pedro Crescêncio Souza. Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy. **GCB Bioenergy**, v. 10, n. 1, p. 52–60, 1 jan. 2018.

SHAKOOR, Nadia; LEE, Scott; MOCKLER, Todd C. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. **Current Opinion in Plant Biology**, v. 38, p. 184–192, 1 ago. 2017.

SHI, Y. et al. Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. **PLoS ONE**, v. 11, n. 7, 1 jul. 2016.

SINGH, Arti et al. Challenges and Opportunities in Machine-Augmented Plant Stress Phenotyping. **Trends in Plant Science**, v. 26, n. 1, p. 53–69, 1 jan. 2021.

TREICHEL, M. **Anuário Brasileiro de Tomate**. Santa Cruz: Editora Gazeta, 2016. Disponível em: <https://www.editoragazeta.com.br/flip/anuario-tomate-2016/files/assets/common/downloads/publication.pdf>. Acesso em: 01 fev. 2025.

VAN DIJK, A. D. J. et al. Machine learning in plant science and plant breeding. **iScience**, v. 24, n. 101890, p. 1–12, 2021.

VELAZCO, J. G. et al. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. **Theoretical and Applied Genetics**, v. 130, n. 7, p. 1375–1392, 1 jul. 2017.

VU, Tien Van et al. Precision Genome Engineering for the Breeding of Tomatoes: Recent Progress and Future Perspectives. **Frontiers in Genome Editing**, v. 2, n. 612137, 2020.

ZSÖGÖN, Agustin et al. De novo domestication of wild tomato using genome editing. **Nature Biotechnology**, v. 36, n. 12, p. 1211–1216, 1 dez. 2018.

4. CAPÍTULO 1:

SELEÇÃO DE GENÓTIPOS DE TOMATEIRO PARA PRODUÇÃO E QUALIDADE DE FRUTOS VIA ÍNDICE FAI-BLUP EMPREGANDO-SE MODELOS MISTOS COM CORREÇÃO ESPACIAL

Felipe de Oliveira Dias, João Marcos Amario de Sousa, Kaio Olimpio das Graças Dias, Carlos Nick Gomes

RESUMO: Comparou-se neste trabalho a implementação do índice FAI-BLUP em dois cenários. No Cenário I, realizou-se o ajuste de diferentes modelos estatísticos, incluindo modelos espaciais, para selecionar aquele de melhor ajuste para cada uma das características avaliadas em tomateiro, estimar os valores BLUP e implementar o índice. No Cenário II, considerou-se apenas o modelo tradicional associado ao delineamento experimental para todas as características e implementação do índice. Em campo, dezesseis híbridos de tomate foram avaliados no delineamento em blocos casualizados (DBC) com quatro repetições. Todas as práticas culturais recomendadas para o cultivo do tomateiro foram adotadas. As características avaliadas foram: produção total de frutos (Prod. total), número de frutos total (Nº frutos), brilho (*L*), tonalidade entre verde e vermelho (*a*), tonalidade entre amarelo e azul (*b*), índice de cromaticidade (*c*), ângulo Hue (*h*), firmeza dos frutos (FF), sólidos solúveis total (SST), pH, acidez titulável (AT), relação SST/AT. No Cenário I, para cada característica foi implementado três modelos: o modelo 1 foi o modelo tradicional de análise de experimentos em DBC, o modelo 2 foi o modelo espacial com resíduos autorregressivos de primeira ordem – AR1 (erro dependente), e o modelo 3 foi o modelo com os resíduos AR1 mais o efeito nugget (erro independente). Nesse caso, para seleção do modelo de melhor ajuste foram utilizados os critérios de AIC, BIC e acurácia. Já no Cenário II, apenas o modelo tradicional foi utilizado na análise de todas as características. Houve variabilidade genética apenas para as características Prod. total, Nº frutos, *a*, *c* e FF. Para o Cenário I, o modelo 1 foi o melhor ajustado para as características Nº frutos, *a*, *c* e FF, e o modelo 3 foi o melhor para Prod. total. Os genótipos foram ranqueados pelo índice FAI-BLUP considerando os dois cenários e observou-se mudança no ordenamento dos genótipos nos diferentes cenários. Considerando uma taxa de seleção de 30% dos genótipos, houve ganho de seleção negativo apenas para a Prod. total pelo índice FAI-BLUP no Cenário II, em direção

contrária ao ideótipo delineado. O FAI-BLUP no Cenário I garantiu ganhos positivos para todas as características, na direção do ideótipo de tomate delineado.

Palavras-chave: *Solanum lycopersicum*, melhoramento genético, acurácia seletiva

4.1. INTRODUÇÃO

A seleção de genótipos superiores é uma tarefa rotineira nos programas de melhoramento genético. Para a seleção simultânea de diversas características em um mesmo genótipo, os melhoristas adotam os índices de seleção. Rocha, Machado e Carneiro (2018), propuseram o índice FAI-BLUP baseado em análise de fatores e desenho do ideótipo. Este índice não apresenta problema de multicolinearidade e nem necessita da atribuição de peso para cada característica considerada, problemas normalmente encontrados no índice clássico Smith-Hazel e nos índices derivados deste último. O FAI-BLUP vem sendo utilizado com ganhos genéticos mais balanceados e com bom desempenho em diversas culturas como em café (BOTEGA, 2019), em trigo (MEIER et al., 2021), em abóbora (OLIVEIRA et al., 2021), em manga (COSTA et al., 2023) e em tomate (COPATI et al., 2024).

Para implementar o índice de seleção FAI-BLUP, os valores genéticos devem ser obtidos para cada característica as quais se deseja obter ganho genético conforme o ideótipo delineado (ROCHA; MACHADO; CARNEIRO, 2018). A predição dos valores genéticos é um ponto central nos programas de melhoramento para o desenvolvimento de cultivares (PIEPHO et al., 2008). Estes valores são obtidos rotineiramente por melhor predição linear não viesada (BLUP) pelo método da máxima verossimilhança restrita (REML) empregando-se modelos mistos (HENDERSON, 1975; PATTERSON; THOMPSON, 1971). Modelos que consideram as tendências espaciais do campo experimental podem melhorar as predições dos valores BLUP (PIEPHO et al., 2008). A metodologia de análise de modelos mistos possibilita a análise estatística espacial, podendo envolver diferentes situações (RESENDE; STURION, 2001). Isso se deve ao fato de ser possível englobar as tendências espaciais na equação geral de modelos mistos pela incorporação de efeitos fixos e aleatórios de acordo com as coordenadas das parcelas no campo experimental (GILMOUR; CULLIS; VERBYLA, 1997).

A análise estatística espacial vem sendo utilizada em muitos trabalhos de melhoramento genético de diversas culturas. Estes trabalhos visam melhorar a

acurácia experimental para melhor predizer os valores genéticos dos genótipos sob seleção, melhorando assim a eficiência seletiva (ANDRADE et al., 2020; COPATI et al., 2021; SALVADOR et al., 2022; SILVA et al., 2024). Tal análise se justifica devido as diversas variações ambientais, não controláveis por delineamento, que podem afetar as unidades experimentais em campo de modo diferente, tornando a pressuposição de erros independentes pouco confiável quando se utiliza apenas o modelo de análise tradicional que assume essa premissa (BURGUEÑO, 2018; RESENDE; STURION, 2001).

Não existe um modelo universal que pode ser aplicado a todo experimento de uma cultura em determinado campo, uma vez que além das variações ambientais naturais no tempo e espaço, variações locais podem ser introduzidas ao longo da condução do experimento (BURGUEÑO, 2018; GILMOUR; CULLIS; VERBYLA, 1997). E distintas características avaliadas em um mesmo experimento podem ser afetadas de modo diferente pelos padrões espaciais do campo experimental (BURGUEÑO, 2018).

Desta forma, o desempenho do índice FAI-BLUP pode ser afetado em relação a metodologia utilizada na predição dos valores genéticos. Assim, esse trabalho buscou implementar o índice FAI-BLUP em dois cenários: Cenário I, que corresponde ao ajuste de modelos estatísticos com resíduos espacialmente dependentes e independentes em diversas características de genótipos de tomateiro, a fim de selecionar o modelo de melhor ajuste para predizer os valores genéticos para cada característica para implementar o índice; Cenário II que corresponde a adoção de um modelo único, conforme o delineamento experimental, para predizer os valores genéticos de todas as características para implementação do índice. O FAI-BLUP implementado em cada cenário foi comparado buscando identificar aquele que possibilitou melhor desempenho do índice na seleção de genótipos de tomateiro.

4.2. MATERIAL E MÉTODOS

4.2.1. Recurso fitogenéticos

Quinze híbridos de tomate foram avaliados quanto a produção e qualidade do fruto. Estes foram obtidos artificialmente com cruzamentos controlados envolvendo 10 genitores (Tabela 1) em casa de vegetação, segundo metodologia apresentada em Nick e Silva (2016) com poucas modificações.

Tabela 1: Características dos genitores utilizados nos cruzamentos para geração dos híbridos de tomateiro.

Genitor	Referência / Empresa	Características dos genitores	Híbridos gerados
NC 25P	Gardner e Panthee (2010a)	Linhagem determinada, gene *Ve* (resistente à Vd), genes *I* e *I-2* (resistente à Fus12), gene *Ph-3* (resistente a Pi) e resistente à As.	1, 2 e 3
Colono®	Sakata	Híbrido determinado, alto nível de resistência à Vd, Fus123, ToMV, Mi, Mj, Ss, As.	1, 13 e 14
Invicto®	Blueseeds	Híbrido determinado, longa vida, tolerância à Vd, Fus12, TMV e TSWV.	2, 4 e 15
87-MC-DF21	Copati et al. (2021)	Geração $F_{3:5}$, resistente à requeima.	3, 5 e 7
NC 2 CELBR	Gardner e Panthee, (2010b)	Linhagem determinada, gene *Ve* (resistente à Vd), genes *I* e *I-2* (resistente à Fus12), genes *Ph-2* e *Ph-3* (resistente a Pi) e resistente à As.	9, 11 e 13
Tainara®	Feltrin	Híbrido determinado, resistência à Vd, TSRV, TMV, Fus123, Mi.	12, 14 e 15
BS DS0005®	Blueseeds	Híbrido determinado, longa vida, tolerâncias: Vd, Fus123, Mi, Mj, ToMV, TSWV, TYLCV e Oi.	4, 5, 6
IL 3-2	Eshed e Zamir (1995)	Linhagem de introgressão de *S. pennellii,* gene *r (yellow fruit flesh)*	7, 8 e 9
IL 2-1	Eshed e Zamir (1995)	Linhagem de introgressão de *S. pennellii.*	10, 11 e 12
123-MC-DF21	Copati et al. (2021)	Geração $F_{3:5}$, resistente à requeima.	6, 8 e 10

As: *Alternaria solani*; Fus12: Fusarium raças 1 e 2; Fus123: *Fusarium* raças 1, 2 e 3; *Mi*: *Meloidogyne incognita*: Mj: *Meloidogyne javanica*; Oi*: Oidium*; Pi: *Phytophthora infestans*; Ss: *Stemphylium solani*; TMV: Tobacco mosaic vírus; ToMV: Tomato mosaic vírus; TSRV: Tomato severe rugose vírus; TSWV: Vírus do Vira-Cabeça; TYLCV: Geminivírus; Vd: *Verticillium dahlia*.

Resumidamente, a metodologia de hibridação consistiu em escolher flores com estigma receptivo, normalmente dois a três dias antes da antese; nestas flores removeu-se duas a três sépalas, todas as pétalas e realizou-se a emasculação; o pólen de outro genitor foi coletado com a flor totalmente aberta e levado até o estigma

da flor emasculada; identificou-se a flor polinizada artificialmente e protegeu-a com um invólucro de papel alumínio, que foi removido naturalmente com o desenvolvimento do fruto; removeu-se ainda as demais flores do cacho (NICK; SILVA, 2016). Além destes, o híbrido comercial Tomate Fascínio® (Empresa Feltrin) foi utilizado como testemunha no experimento, totalizando assim 16 tratamentos.

4.2.2. Implantação e condução do experimento

O experimento foi implantado e conduzido no município de Viçosa – MG, localizado a 648,74 m de altitude e coordenadas geográficas 20º 45’ 14’’ S e 42º 52’ 53” W. O clima regional, pela classificação de Köppen, é do tipo Cwb com verões chuvosos e invernos secos. A área experimental pertence a Unidade de Ensino, Pesquisa e Extensão “Horta Velha” do Departamento de Agronomia da Universidade Federal de Viçosa. Desta área foi obtida a amostra composta de solo da camada superficial de 0 – 20 cm que foi utilizada para realizar a análise química. Com o resultado e realizada a interpretação da análise de solo, foi feita a recomendação de calagem e adubação conforme a 5ª aproximação para a cultura do tomateiro (FILGUEIRA et al.,1999). O solo foi arado e gradeado em toda área experimental e sulcado nas linhas de plantio das mudas com os implementos agrícolas apropriados. O calcário recomendado foi aplicado nas linhas de plantio e incorporado manualmente com enxadas.

Os tratamentos foram semeados em bandejas de polietileno de 64 células com volume de 30 cm³, preenchidas com substrato comercial. As bandejas ficaram em ambiente protegido para o desenvolvimento das mudas. A irrigação ocorreu diariamente de acordo com a necessidade. A fertirrigação foi conforme a recomendação proposta por Furlani et al. (2013). As mudas estavam prontas para o transplantio quando apresentavam de 4 – 5 folhas definitivas expandidas por completo. A desbrota dos tomateiros foi realizada até o primeiro cacho. O tutoramento foi feito com uso de fitilho na horizontal, sustentado por tutores distribuídos a cada duas plantas na linha de plantio. Houve o parcelamento da adubação recomendada conforme proposição de Alvarenga et al. (2013).

4.2.3. Croqui experimental

Foi utilizado o delineamento em blocos casualizados (DBC) de modo que os blocos controlavam o suave declive do terreno. Cinco plantas constituíam cada parcela

experimental. No total, foram utilizados quatro blocos e 64 parcelas experimentais. O posicionamento das parcelas no campo, de acordo com as coordenadas de linha e coluna, foi de acordo com a Figura 1. As linhas variaram de 1 a 4 enquanto as colunas variaram de 1 a 16. A cada quatro colunas um bloco é formado. O espaçamento adotado foi de 0,60 x 1,20 m entre plantas e entre linhas, respectivamente.

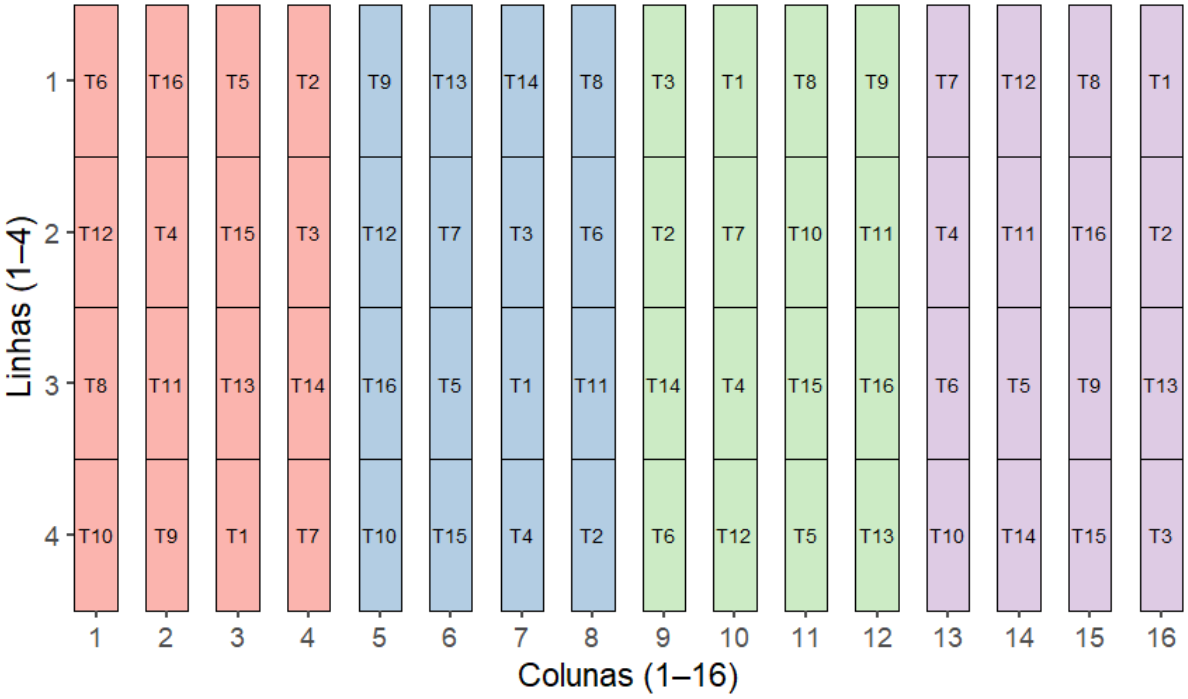

Figura 1: Croqui experimental dos 16 genótipos de tomateiro avaliados em DBC com as coordenadas de linha e coluna das parcelas no campo. T1 a T15 identificaram os híbridos, T16 identificou a testemunha.

4.2.4. Características avaliadas

Os genótipos do tomateiro foram avaliados quanto a produção total, em kg (Prod. total). Para isso, semanalmente os frutos no estádio verde-maduro foram colhidos e pesados em balança. Também, houve a contagem de cada unidade para obter o número de frutos total (Nº frutos). Para as características de qualidade, uma amostra de cinco frutos do terço mediano das plantas foi avaliada no Laboratório de Manejo de Recursos Genéticos do Departamento de Agronomia. Com uso de um colorímetro, avaliou-se a coloração dos frutos na região equatorial. Ajustou-se o colorímetro para o sistema de leitura *L a b* e a calibração foi feita em placa branca padrão. As medidas de *L* indicam o brilho/luminosidade, variando de 0 a 100, onde 0 indica falta de brilho e 100 corresponde ao brilho máximo. Já as medidas de *a* indicam a tonalidade, variando de -60 a +60, onde valores negativos equivalem ao verde e os positivos

equivalem ao vermelho. As medidas de *b* indicam outra escala de tonalidade, também variando de -60 a +60, onde valores negativos equivalem ao amarelo e positivos equivalem ao azul. As medidas do índice de cromaticidade *c* indicam a saturação ou pureza da cor. As medidas do ângulo Hue (*h*) indicam o grau da cor (0º = vermelho; 90º = amarelo; 180º = verde; 270º = azul). Com uso de um penetrômetro avaliou-se a firmeza dos frutos (FF). As medidas foram tomadas na região equatorial dos frutos, expressas em Newton (N), indicando a força necessária para romper o pericarpo dos frutos. Com auxílio de um pHmetro avaliou-se o potencial hidrogeniônico (pH) do suco integral de tomate, ocorrendo a calibração com solução tampão de pH 4 e 7. Com um refratômetro digital foi obtido o teor de sólidos solúveis total (SST) expresso em ºBrix. Para obter a acidez titulável (AT) da polpa homogeneizada, uma amostra de 10 g foi diluída em 50 ml de água destilada e titulada com solução padrão de hidróxido de sódio a 0,1N e fenolftaleína como indicador. Dividindo os valores de sólidos solúveis pela acidez titulável foi obtido a relação SST/AT.

4.2.5. Análises estatísticas

4.2.5.1. Modelos estatísticos ajustados

Consideramos o seguinte modelo para a análise individual:

$$y = \mu 1 + Xb + Zg + \epsilon$$

Onde: *y* é o vetor n × 1 de observações fenotípicas;

b é o vetor r × 1 de efeitos fixos de blocos;

g é o vetor m × 1 de efeitos aleatórios de genótipos, assumindo $g \sim N(0, \sigma_g^2)$, onde σ_g^2 é a variância genotípica;

ϵ é o vetor n × 1 de efeitos residuais aleatórios, assumindo $\epsilon \sim N(0, \sigma_\epsilon^2)$, onde σ_ϵ^2 é a variância residual;

X e *Z* são as matrizes de incidência para esses efeitos.

Ajustamos três estruturas de covariância residual para modelar os efeitos residuais: o modelo 1 (MOD1) não considerou as correlações entre linhas ou colunas, ou seja, $\epsilon \sim N(0, \sigma_\xi^2 I_r \otimes I_c)$, onde σ_ξ^2 é a variância residual, e I_r e I_c são matrizes identidade que representam independência espacial nas direções de linha e coluna, respectivamente. O modelo 2 (MOD2) considerou correlações em ambas as direções (linhas e colunas), ou seja, $\epsilon \sim N(0, \sigma_\xi^2 \Sigma_r(\rho_r) \otimes \Sigma_c(\rho_c))$, onde $\Sigma_c(\rho_c)$ é a matriz de

correlação autorregressiva de primeira ordem para as colunas; $\Sigma_r(\rho_r)$ é a matriz de correlação autorregressiva de primeira ordem para as linhas; σ_ξ^2 é a variância dos resíduos espacialmente correlacionados e $\otimes$ é o produto de Kronecker. O modelo 3 (MOD3) também considerou correlações em ambas as direções (linhas e colunas) por meio de uma estrutura autorregressiva de primeira ordem, porém incluiu variação adicional independente entre as unidades experimentais. Assim, temos: $\epsilon \sim N(0, \sigma_\xi^2 \Sigma_r(\rho_r) \otimes \Sigma_c(\rho_c))$, onde $\Sigma_r(\rho_r)$ é a matriz de correlação autorregressiva de primeira ordem para as linhas e $\Sigma_c(\rho_c)$ é a matriz de correlação autorregressiva de primeira ordem para as colunas. A variação adicional foi modelada usando a estrutura idv(units), que assume que as unidades experimentais possuem variações independentes entre si. Pode também ser chamado de efeito nugget ou pepita, representando o erro de mensuração de cada parcela ou as tendências locais.

No Cenário I, para cada característica avaliada ajustou-se os três modelos mistos, enquanto no Cenário II apenas o modelo 1 foi considerado para todas as características. Os BLUPs de cada característica em cada genótipo foram preditos de acordo com o modelo selecionado.

4.2.5.2. Seleção de modelos

Para auxiliar a escolha do modelo melhor ajustado a cada característica, no caso do Cenário I, alguns critérios foram utilizados:

- **Variabilidade genética**

O intervalo de confiança (IC) para a variabilidade genética de cada modelo ajustado para cada característica foi gerado com nível de confiança de 95%, do seguinte modo:

$$\mathrm{IC}(\sigma_g^2)_{0,95}: \hat{\sigma}_g^2 \pm 1{,}96\mathrm{xEP}$$

Onde: $\hat{\sigma}_g^2$ é a variância genética estimada;

EP é o erro padrão da variância genética estimada.

Sob as hipóteses H_0: $\sigma_g = 0$ e H_a: $\sigma_g \neq 0$, concluiu-se pela não rejeição de H_0 se o valor testado (zero) para a variância genética estiver contido dentro do intervalo. Caso contrário, rejeitou-se a hipótese nula.

- **AIC e BIC**

Os critérios de informação de Akaike – AIC (AKAIKE, 1974) e bayesiano – BIC (SCHWARZ, 1978) indicam o melhor modelo de ajuste por seus menores valores. São calculados da seguinte forma:

$$AIC = -2logL + 2p$$

$$BIC = -2logL + plog[n - r(x)]$$

Onde: *L* é o valor que torna máxima a função de verossimilhança do modelo;

p é o número de parâmetros estimados no modelo;

n é o número de observações;

r(x) é o posto da matriz de incidência dos efeitos fixos.

- **Herdabilidade**

A herdabilidade no sentido amplo (H_g^2) de cada característica em cada modelo foi estimada conforme segue (CULLIS; SMITH; COOMBES, 2006):

$$H_g^2 = 1 - \frac{\bar{v}BLUP}{2\hat{\sigma}_g^2}$$

Onde: $\bar{v}BLUP$ é a variância média das diferenças entre pares de BLUPs dos efeitos genéticos;

$\hat{\sigma}_g^2$ é a variância genética estimada.

- **Acurácia**

A acurácia ($r_{\hat{g}g}$) dos modelos foi estimada conforme se segue:

$$r_{\hat{g}g} = \sqrt{1 - \frac{PEV}{\hat{\sigma}_g^2}}$$

Onde: *PEV* é a variância do erro de predição;

$\hat{\sigma}_g^2$ é a variância genética estimada.

4.2.5.3. Implementação do índice FAI-BLUP

O índice FAI-BLUP foi implementado considerando dois cenários: o Cenário I considerou o estudo e a seleção do melhor modelo para predição dos valores

genéticos de cada característica (entre os modelos 1, 2 e 3); e o Cenário II considerou um único modelo baseado no delineamento experimental para predição dos valores genéticos de todas as características (modelo 1 apenas).

Os valores genéticos preditos via BLUP para cada característica foram submetidos ao índice de seleção FAI-BLUP, conforme Rocha, Machado e Carneiro (2018). Este índice utiliza a análise de componentes principais para extrair as cargas fatoriais da matriz de correlação dos valores genéticos. Para a rotação analítica e o cálculo dos escores fatorais do método dos mínimos quadrados ponderados (BARTLETT, 1978) foi utilizado o critério varimax (KAISER, 1958). O ideótipo para implementar o FAI-BLUP foi construído conforme a Tabela 2, baseado no comportamento (aumentar ou diminuir) que se espera para cada característica que teve variabilidade genética significativa.

Tabela 2: Ideótipo delineado para implementação do índice FAI-BLUP em tomateiro para as características com variabilidade genética significativa.

Ideótipo	**Prod**	**Frutos**	***a***	***c***	**FF**
Desejável	max	max	max	max	max
Indesejável	min	min	min	min	min

Prod: produção total de frutos (kg); Frutos: nº de frutos total; *a*: tonalidade entre verde e vermelho; *c*: índice de cromaticidade; FF: firmeza dos frutos (N); max : BLUP máximo da característica; min: BLUP mínimo da característica.

As distâncias entre genótipo e ideótipo são estimadas e convertidas em probabilidade espacial, possibilitando o ranqueamento dos genótipos para posterior seleção. O algoritmo utilizado foi o seguinte:

$$P_{ij} = \frac{\frac{1}{d_{ij}}}{\sum_{i=1;j=1}^{i=n;j=m} \frac{1}{d_{ij}}}$$

Onde: P_{ij} é a probabilidade do i-ésimo genótipo (i = 1, 2, ..., n) ser semelhante ao j-ésimo ideótipo (j = 1, 2, ..., m);

d_{ij} é a distância genótipo-ideótipo do i-ésimo genótipo ao j-ésimo ideótipo com base na distância euclidiana média padronizada.

4.2.5.4. Ganho com a seleção

A predição do ganho genético com a seleção (GS) dos genótipos foi feita da seguinte maneira:

$$GS_c(\%) = \frac{\overline{x_s} - \overline{x_o}}{\overline{x_o}} \, x \, 100$$

Onde: c é a característica considerada;

$\overline{x_s}$ é a média dos valores genéticos dos genótipos selecionados para c;

$\overline{x_o}$ é a média dos valores genéticos de todos os genótipos para c.

4.2.5.5. Software

A análise dos dados foi feita no Software R (R Core Team 2024) empregando-se a metodologia de modelos mistos. Os modelos foram implantados utilizando o pacote ASRemlR versão 4.2 (BUTLER et al., 2023) que estima os componentes de variância dos modelos por máxima verossimilhança residual (REML). O índice FAI-BLUP foi implementado com a rotina de análise do pacote metan (OLIVOTO; LÚCIO, 2020).

4.3. RESULTADOS

Para o Cenário I, a Tabela 3 apresenta as estimativas dos parâmetros dos modelos selecionados para cada característica, bem como os critérios utilizados em sua seleção. O modelo 3 não atingiu a convergência para número de frutos total (Nº frutos), brilho (*L*), tonalidade entre o verde e o vermelho (*a*), tonalidade entre o amarelo e o azul (*b*), ângulo Hue (*H*), firmeza do fruto (FF), e sólidos solúveis total (SST).

O intervalo de confiança com nível de 95% de confiança ($\mathrm{IC}(\sigma_g^2)_{0,95}$) para a variância genética demonstrou que apenas as características produção total de frutos (Prod. total), Nº frutos, *a*, *c* e FF possuem variabilidade genética diferente de zero. Como a variância genética é importante para se obter ganhos genéticos significativos com a seleção, as demais características não foram consideradas para as análises seguintes (ROCHA; MACHADO; CARNEIRO, 2018).

Tabela 3: Modelos ajustados para cada característica avaliada em tomateiro e seus respectivos critérios para seleção do modelo de melhor ajuste (Cenário I). Para as características com variabilidade genética significativa os modelos selecionados estão em negrito.

Variável	Modelo	$\widehat{\sigma}_g^2$	$\widehat{\sigma}_e^2$	$\widehat{\sigma}_\eta^2$	$\widehat{\sigma}_\xi^2$	ρ_{lin}	ρ_{col}	H_g^2	AIC	BIC	$r_{\hat{g}g}$
	1	7,31	8,43	-	-	-	-	0,78	225,46	229,65	0,88
Prod. total*	2	8,19	-	-	10,13	0,42	0,36	0,84	224,39	232,77	0,88
	3	**7,17**	**-**	**4,19**	**94,63**	**0,98**	**0,98**	**0,84**	**214,86**	**221,15**	**0,88**
	1	**1775,70**	**1539,62**	**-**	**-**	**-**	**-**	**0,82**	**541,31**	**545,49**	**0,91**
Nº frutos*	2	1690,92	-	-	1574,19	-0,01	0,10	0,82	544,94	553,32	0,90
	3 nc	-	-	-	-	-	-	-	-	-	-
	1	0,45	2,74	-	-	-	-	0,40	143,19	147,38	0,58
*L*ns	2	0,39	-	-	2,73	-0,22	0,01	0,37	145,45	153,83	0,58
	3 nc	-	-	-	-	-	--	-	-	-	-
	1	**3,61**	**3,74**	**-**	**-**	**-**	**-**	**0,79**	**177,99**	**182,18**	**0,89**
*a**	2	3,37	-	-	3,89	0,10	0,08	0,78	181,57	189,95	0,88
	3 nc	-	-	-	-	-	-	-	-	-	-
	1	1,47	2,89		-	-	-	0,67	155,33	159,51	0,81
*b*ns	2	1,52	-	-	2,89	-0,02	0,09	0,68	159,07	167,45	0,82
	3 nc	-	-	-	-	-	-	-	-	-	-
	1	**3,48**	**3,34**	**-**	**-**	**-**	**-**	**0,81**	**172,12**	**176,30**	**0,90**
*c**	2	4,16	-	-	4,35	0,46	0,44	0,88	171,07	179,45	0,90
	3	4,17	-	0,00	4,37	0,46	0,44	0,88	173,07	183,54	0,90
	1	1,74	2,87		-	-	-	0,71	156,74	160,93	0,84
*h*ns	2	1,77	-	-	2,85	-0,17	0,07	0,72	160,02	168,39	0,85
	3 nc	-	-	-	-	-	-	-	-	-	-
FF*	**1**	**17,54**	**11,59**	**-**	**-**	**-**	**-**	**0,86**	**251,37**	**255,55**	**0,93**

	2	17,62	-	-	11,69	0,08	0,01	0,86	255,14	263,52	0,93
	3 nc	-	-	-	-	-	-	-	-	-	-
	1	0,0692	0,04		-	-	-	0,87	-84,96	-80,77	0,93
SST^{ns}	2	0,0745	-	-	0,04	-0,33	-0,02	0,89	-83,50	-75,12	0,95
	3 nc	-	-	-	-	-	-	-	-	-	-
	1	0, 0015	0, 01		-	-	-	0,60	-243,69	-239,50	0,77
AT^{ns}	2	0,0016	-	-	0,0038	-0,17	0,01	0,62	-240,42	-232,04	0,79
	3	0,0016	-	0,00	0,0038	-0,17	0,01	0,63	-238,42	-227,95	0,79
	1	0,3371	0,47		-	-	-	0,74	50,26	54,45	0,86
SST/AT^{ns}	2	0,3300	-	-	0,4816	0,12	0,01	0,74	53,89	62,27	0,85
	3	0,3298	-	0,00	0,4813	0,12	0,01	0,74	55,89	66,36	0,85
	1	0,0013	0,01		-	-	-	0,58	-246,43	-242,24	0,75
pH^{ns}	2	0,0013	-	-	0,0038	-0,03	0,02	0,58	-242,45	-234,08	0,75
	3	0,0013	-	0,00	0,0038	-0,03	0,02	0,58	-240,45	-229,98	0,75

* : significativo pelo $\mathrm{IC}(\sigma_g^2)_{0,95}$; ^{ns} : não significativo pelo $\mathrm{IC}(\sigma_g^2)_{0,95}$; Prod. total: produção total (kg); Nº frutos: número de frutos total; *L* : brilho; *a*: tonalidade entre verde e vermelho; *b* : tonalidade entre amarelo e azul; c: índice de cromaticidade; *h* : ângulo Hue; FF: firmeza do fruto; SST : sólidos solúveis total; AT : acidez titulável; SST/AT : relação entre sólidos solúveis total e acidez titulável; pH : potencial hidrogeniônico; ^{nc} : modelo não atingiu convergência; $\hat{\sigma}_g^2$: estimativa da variância genética; $\hat{\sigma}_e^2$: estimativa da variância do erro espacialmente independente; $\hat{\sigma}_\eta^2$: estimativa da variância do efeito nugget; $\hat{\sigma}_\xi^2$: estimativa da variância do erro espacialmente dependente; ρ_{lin}: coeficiente de autocorrelação na linha; ρ_{col}: coeficiente de autocorrelação na coluna; H_g^2: herdabilidade; AIC: critério de informação de Akaike; BIC: critério de informação bayesiano; $r_{\hat{g}g}$: acurácia.

Os critérios de informação de Akaike (AIC) e bayesiano (BIC) indicam o modelo de melhor ajuste por seus menores valores. O modelo 3 foi selecionado para Prod. total, com menores valores de AIC e BIC e maiores valores de herdabilidade e acurácia. De modo semelhante, o modelo 1 foi selecionado para Nº frutos, *a* e FF, com menores valores de AIC e BIC e maiores valores de herdabilidade e acurácia. Já para *c* o modelo que apresentou menor valor de AIC foi o modelo 2. Entretanto, pelo critério bayesiano o modelo 1 foi que apresentou menor valor. A diferença entre os valores de AIC de dois modelos deve ser maior do que duas unidades para se considerar o modelo de menor valor como mais bem ajustado do que aquele de maior valor, sendo ambos os modelos classificados no mesmo patamar de qualidade de ajuste se essa diferença for menor (CAVANAUGH; NEATH, 2019). O mesmo raciocínio pode ser aplicado a diferença entre valores de BIC de dois modelos (NEATH; CAVANAUGH, 2012). Para *c*, a diferença entre os valores de AIC entre os modelos 1 e 2 é de 1,05. Já para BIC essa diferença é de 3,15. Logo, podemos afirmar que pelo critério de AIC os modelos 1 e 2 possuem a mesma qualidade de ajuste, enquanto para o critério bayesiano o modelo 1 se ajustou melhor. A herdabilidade desse modelo foi menor do que dos outros, entretanto a acurácia não diferiu entre eles. Por isso o modelo 1 também foi selecionado para a predição dos valores genéticos para *c*.

Apenas para a Prod. total (kg) houve a seleção de um modelo com erro espacialmente dependente (modelo 3). Nesse modelo, o efeito do erro aleatório nugget e do efeito do erro correlacionado foi de 4,19 e 94,63 respectivamente. O coeficiente de autocorrelação espacial foi de 0,98 tanto no sentido das linhas quanto no sentido das colunas.

Os valores das médias BLUP estimados para cada característica de cada genótipo, considerando a predição pelo modelo selecionado, estão apresentados na Tabela 4.

Tabela 4: Valores das médias BLUP das características com variabilidade genética avaliadas em tomateiro.

Genótipo	Prod	Frutos	*a*	*c*	FF
T1	21,57	181,84	37,92	57,68	19,15
T2	21,82	191,09	34,47	56,85	21,40
T3	21,05	219,44	37,95	57,93	17,11
T4	20,70	126,37	38,63	59,25	25,71
T5	20,19	144,24	40,35	61,76	27,05
T6	19,94	146,09	39,11	61,11	25,28
T7	16,41	188,42	39,45	60,49	22,97
T8	16,96	200,95	40,79	62,41	21,27
T9	14,65	147,53	39,46	60,60	21,83
T10	17,87	216,98	41,34	61,36	22,35
T11	16,98	196,43	41,17	61,59	23,62
T12	18,14	229,10	40,19	60,70	27,04
T13	17,17	101,71	37,92	57,73	22,79
T14	18,44	119,59	38,97	58,96	27,75
T15	20,77	167,67	39,07	59,99	26,86
T16	23,89	191,29	40,58	59,39	33,22

Prod: produção total de frutos (kg); Frutos: nº de frutos total; *a*: tonalidade entre verde e vermelho; *c*: índice de cromaticidade; FF: firmeza dos frutos (N); T1 a T15 identificaram os híbridos, T16 identificou a testemunha.

A Figura 2 mostra as estimativas de autovalores e a variância acumulada para as cinco componentes principais obtidas pela matriz de correlação genética. O Cenário I é representado pela Figura 2 A e o Cenário II é representado pela Figura 2 B. O número de fatores para condensar os dados é igual ao número de autovalores maior do que um, de acordo com o critério de Kaiser (KAISER, 1958). Apenas três componentes obtiveram autovalores maior do que um em ambos os cenários, indicando a redução da dimensionalidade em três fatores. Essas três componentes explicam aproximadamente 94% e 93% da variância genética para os Cenários I e II, respectivamente.

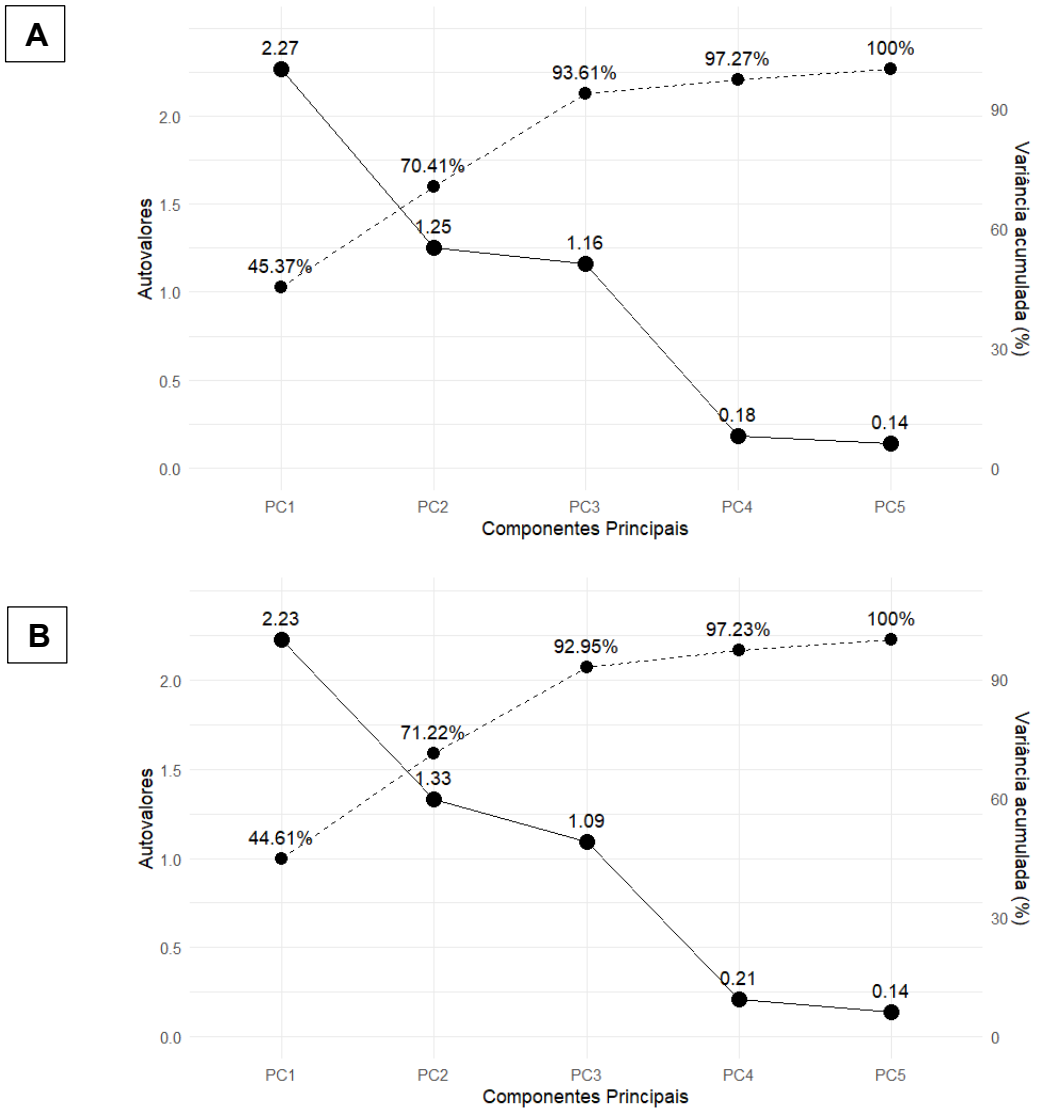

Figura 2: Estimativas de autovalores e variância acumulada das componentes principais para implementar o FAI-BLUP em tomateiro: **A** = Cenário I; **B** = Cenário II.

A Figura 3 apresenta o mapa de calor para as cargas fatoriais após a rotação varimax e a comunalidade. Para o Cenário I, o primeiro fator apresentou alta correlação genética positiva com *a* e *c*, o segundo fator se correlacionou positivamente com FF e Prod. total, enquanto o terceiro fator se correlacionou negativamente com o Nº frutos. A comunalidade entre os fatores e as características variou de 0,91 a 0,97. Já para o Cenário II, o primeiro fator apresentou alta correlação genética positiva com *a*, *c* e negativa com Prod. total, o segundo fator se correlacionou positivamente com FF, enquanto o terceiro fator se correlacionou positivamente com o Nº frutos. A comunalidade entre os fatores e as características variou de 0,91 a 0,96.

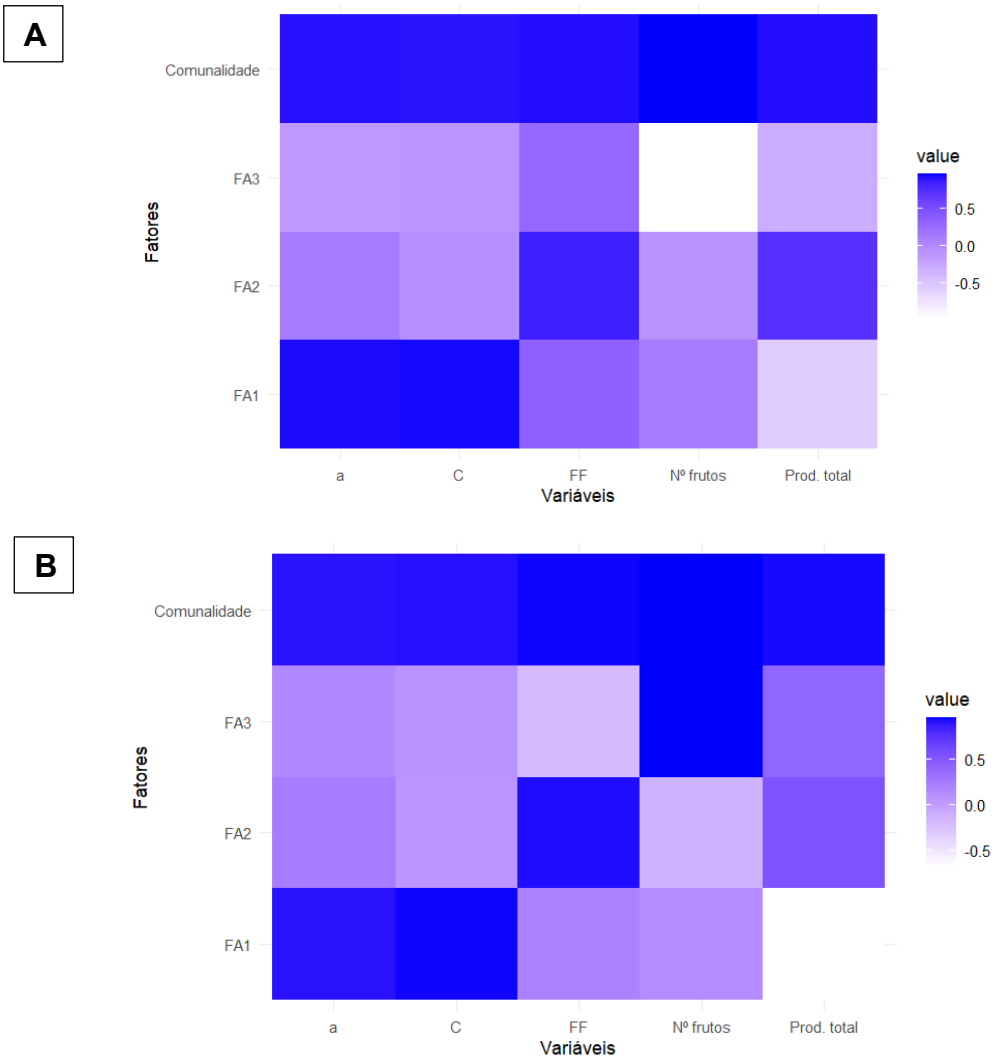

Figura 3: Mapa de calor para as cargas fatoriais após a rotação varimax e comunalidade para implementação do FAI-BLUP: **A** = Cenário I; **B** = Cenário II. *a* = tonalidade entre verde e vermelho; *c* = índice de cromaticidade; FF = firmeza do fruto; Nº frutos = número de frutos total; Prod. total = produção total (kg).

A Figura 4 apresenta o ranqueamento dos genótipos de tomateiro pelo índice FAI-BLUP para os dois cenários. O Cenário I representa a situação de estudo e seleção adequada de modelos para predição dos valores genéticos para as características a serem consideradas no índice. O Cenário II representa a situação que corresponde a metodologia tradicionalmente utilizada de implementação do índice, com modelo único para predição dos valores genéticos para todas as características. Houve mudança de posição dos genótipos no ranque de ambos os cenários, que pode interferir diretamente na seleção a depender da taxa de seleção adotada.

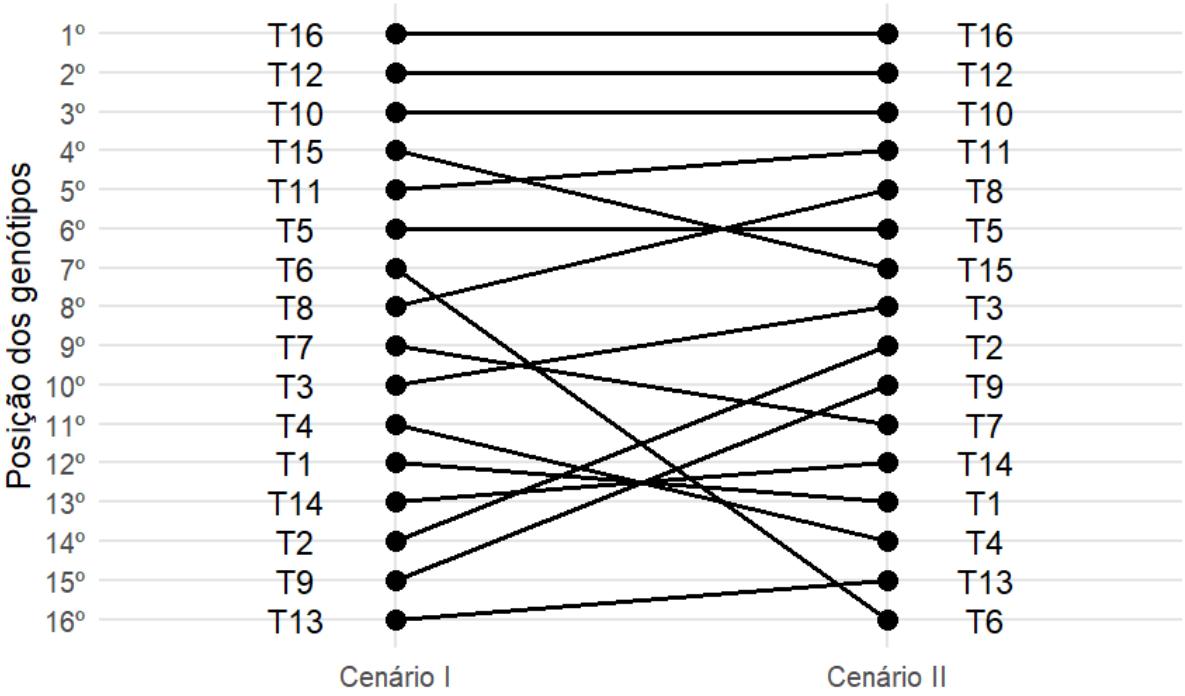

Figura 4: Ranqueamento dos genótipos pelo índice FAI-BLUP para os Cenários I (com análise espacial) e II (sem análise espacial).

Considerando uma taxa de seleção de 30% dos melhores genótipos, a Tabela 5 mostra os resultados da seleção direta para cada característica e da seleção pelo índice FAI-BLUP para ambos os cenários. Conforme ideótipo delineado na Tabela 2, são esperados ganhos positivos para todas as características consideradas. Apenas para a Prod. total houve ganho com a seleção negativo considerando o índice FAI-BLUP no Cenário II. Esta característica foi a única afetada pelas tendências espaciais do campo experimental, conforme Tabela 3.

Tabela 52: Ganho com a seleção (GS) de 30% dos melhores genótipos para a seleção direta da característica e pela seleção com o índice FAI-BLUP para os Cenários I e II.

Variável	X_o	Seleção direta		Cenário I		Cenário II	
		X_s	GS (%)	X_s	GS (%)	X_s	GS (%)
Prod. total	19,16	21,82	13,88	19,53	1,94	18,77	-2,03
Nº frutos	173,05	212,58	22,85	200,29	15,75	206,95	19,59
a	39,21	40,85	4,17	40,47	3,21	40,81	4,09
c	59,86	61,64	2,98	60,61	1,24	61,09	2,05
FF	24,09	28,39	17,84	26,62	10,51	25,50	5,87

Prod. total = produção total de frutos (kg); Nº frutos = número de frutos total; *a* = tonalidade entre o verde e vermelho; *c* = índice de cromaticidade; FF = firmeza de frutos (N); X_o = média dos valores genotípicos do conjunto de genótipos; X_s = média dos valores genotípicos dos genótipos selecionados.

4.4. DISCUSSÃO

Segundo Cavanaugh e Neath (2019) ao se considerar diferentes modelos para descrever o mesmo fenômeno, tem-se o desafio de escolher o modelo que mais se aproxima da realidade, o que pode ser remetido à máxima de George Box “Todos os modelos estão errados, mas alguns são úteis.” Ainda segundo esses autores, os critérios de informação possuem como princípio a rejeição de modelos que são muito simples ou desnecessariamente parametrizados para acomodar os dados. Considerando que o modelo real não seja um dos candidatos, o AIC é um critério eficiente que garante que o modelo selecionado seja o modelo ótimo que minimiza o erro da predição (CAVANAUGH; NEATH, 2019). Já BIC é considerado um critério consistente que garante a seleção do modelo real, se ele estiver presente entre os candidatos, ou quase real, em caso contrário, além de ser considerado mais parcimonioso que AIC (NEATH; CAVANAUGH, 2012). Apenas para *c* houve divergência entre os menores valores de AIC e BIC, e baseado na parcimônia BIC levou a seleção do modelo 1 em detrimento do modelo 2, que foram considerados com a mesma qualidade de ajuste por AIC.

As variações ambientais podem ser controladas pelo delineamento experimental e pela metodologia de análise espacial, possibilitando uma comparação mais eficiente e acurada do efeito genético entre tratamentos (HOEFLER et al., 2020). Este trabalho demonstrou que apenas a blocagem via delineamento experimental não foi suficiente para controlar as tendências espaciais para predizer os valores genéticos para a Prod. total de tomate. Para isso, o modelo 3 que considera a autocorrelação espacial se mostrou ser o mais adequado. Já para as demais características, associadas a qualidade do tomateiro, o delineamento em blocos casualizados foi suficiente para garantir o controle ambiental, uma vez que o modelo 1 se mostrou ser melhor que os demais para predizer os valores BLUP conforme os critérios de informação.

A seleção de diferentes modelos para a predição dos valores genéticos de tomateiros avaliados em um mesmo experimento evidencia que a dependência espacial afeta de maneira diferente as características sob seleção. De modo semelhante, Bernardeli et al. (2021), relataram a escolha de diferentes modelos selecionados para predizer os valores genéticos para três características avaliadas em grãos de soja de ensaios multi-ambiente: teor de proteína, teor de óleo e teor relativo de proteína. Estes autores selecionaram modelos estatísticos que além de se

diferirem para os ambientes considerados, em se tratando do mesmo ambiente, ao menos para uma das características o modelo melhor ajustado foi diferente para as demais.

A magnitude do valor de autocorrelação espacial obtido para o modelo 3 indica que a Prod. total foi afetada por fatores não controlados por meio do delineamento experimental. Valores acima de 0,60 para a autocorrelação demonstram forte dependência espacial, associadas a padrões locais no campo experimental (VELAZCO et al., 2017). Estes padrões podem ser devidos a inúmeros fatores como as variações nas propriedades químicas, físicas e biológicas do solo, variações climáticas e operações do manejo fitotécnico ao longo da condução do experimento (GILMOUR; CULLIS; VERBYLA, 1997). Controlar todos esses fatores é um desafio na experimentação agrícola, sendo mais eficiente o controle quanto menor for a área total do bloco, considerando apenas o delineamento, ou em conjunto com a modelagem das tendências espaciais (BURGUEÑO, 2018).

Além do erro correlacionado, o modelo 3 também engloba o erro aleatório nugget. Quando não se modela esse efeito a variância genética pode ser superestimada como consequência (VELAZCO et al., 2017). Tal fato pode ser observado comparando-se os modelos 2 e 3 para a Prod. total. O modelo 2 que não considera o efeito nugget obteve uma estimativa da variância genética maior que a estimativa da variância genética do modelo 3.

Os altos valores de acurácia obtidos nesse trabalho possibilitam uma seleção mais realística em termos genéticos. A análise de componentes principais para implementação do índice FAI-BLUP é utilizada para extrair as cargas fatoriais da matriz de correlação genética, sendo o número de fatores igual ao número de componentes com autovalores maior que um (ROCHA; MACHADO; CARNEIRO, 2018). Considerando ambos os cenários de implementação do FAI-BLUP, os resultados obtidos pelas componentes principais foram semelhantes.

A comunalidade é a proporção da variação nas características explicadas por todos os fatores extraídos (OLIVEIRA et al., 2021). Altos valores de comunalidade indicam que os fatores explicam boa parte da variância de uma característica, ou seja, os fatores descrevem bem as características, garantido assim a confiabilidade da análise fatorial (BOJARIAN; ASADI-GHARNEH; GOLABADI, 2018). Nesse estudo, a comunalidade foi alta na implementação do FAI-BLUP nos dois cenários considerados.

A correlação genética é considerada em cada fator, garantindo que as relações entre características sejam preservadas, ou seja, características agrupadas no mesmo fator são mais correlacionadas entre si do que aquelas de fatores diferentes (ROCHA; MACHADO; CARNEIRO, 2018). Dessa forma, características agrupadas no mesmo fator podem ser selecionadas indiretamente ao se praticar seleção direta para uma das características do fator, sendo esta prática uma forma eficiente de otimizar recursos (COSTA et al., 2023). A Prod. total mudou do fator 2, no Cenário I, para o fator 1, no Cenário II, com a implementação do FAI-BLUP. No Cenário I, a seleção indireta para aumento dessa característica deveria ser feita selecionando-se para o aumento da FF. Já no Cenário II, a seleção indireta para aumento da Prod. total deveria ser feita pela redução de *a* ou *c*. Dessa forma, a seleção indireta para características afetadas pela dependência espacial pode ser erroneamente realizada ao não se considerar o efeito dessa dependência no modelo de predição dos valores genéticos. Além disso, não seria vantajoso a seleção indireta reduzindo-se uma característica que se deseja aumentar, como ocorreria se se praticasse seleção indireta no fator 1 do Cenário II.

A mudança no posicionamento dos genótipos em cada cenário reforça a necessidade da escolha adequada do modelo para classificação mais acurada dos genótipos pelo índice FAI-BLUP. Costa et al. (2023), utilizaram o índice FAI-BLUP para selecionar genótipos de manga quanto a características de qualidade da fruta colhidas em três safras diferentes. Em seu trabalho, esses autores verificaram uma forte influência ambiental nas características *b* da casca, *L*, *c* e *h* da polpa das frutas que obtiveram a variância ambiental temporária explicando a maior parte da variância fenotípica total. Apesar dessa constatação, o modelo básico para repetibilidade sem delineamento experimental foi utilizado para a predição dos BLUPs para todas as características. Como resultado da seleção dos 25 melhores genótipos de manga, os autores verificaram ganhos negativos para *a* e positivos para *h* que divergiu do ideótipo delineado, ou seja, deveria ter aumento em *a* e redução em *h*. Nesse caso, de acordo com Resende e Sturion (2001), modelar a dependência temporal pode ser vantajoso no caso de medidas repetidas, uma vez que essa dependência aumenta com a diminuição da distância entre as observações. Tal resultado pode ser comparado ao obtido neste trabalho, ao se considerar o Cenário II. A falta de modelagem da dependência espacial resultou na divergência entre o ganho genético esperado e o ganho estimado para Prod. total de tomate. Já com o estudo prévio e

seleção do modelo de melhor ajuste para proceder a predição dos valores BLUP, Cenário I, o ganho genético estimado foi positivo conforme desejado.

As mudanças no ranqueamento podem ser ainda mais significativas ao se considerar o aumento do número de genótipos, levando também ao aumento na área experimental que poderá acarretar o aumento da dependência espacial. Essa mudança também foi relatada no trabalho de Salvador et al. (2022) em experimentos de melhoramento genético de feijão comum. Os autores verificaram a autocorrelação espacial afetando a comparação dos genótipos, que foi melhor controlada pela modelagem das tendências espaciais, resultando em maior acurácia da predição dos valores genéticos. Isso reforça ainda a necessidade de utilizar critérios de informação para seleção do modelo de melhor ajuste, uma vez que mudanças no posicionamento dos genótipos são esperadas ao se considerar diferentes modelos, e os critérios auxiliam na seleção mais precisa de genótipos superiores (SILVA et al., 2024).

A mudança no ranqueamento dos genótipos pode ainda afetar a eficiência da seleção do FAI-BLUP considerando diferentes taxas de seleção. Copati et al. (2021) selecionaram um modelo com resíduo espacialmente dependente como o mais adequado para predizer os valores genéticos de 200 famílias de tomateiro quanto a resistência a requeima. Na seleção direta e comparando o modelo espacial selecionado com o modelo tradicional do delineamento utilizado, os autores verificaram que para a taxa de seleção de 10% a eficiência seletiva foi de 35%, enquanto para a taxa de 20% a eficiência foi de 50%. Nesse estudo, o FAI-BLUP para os dois cenários começa a divergir entre os genótipos selecionados a partir da quarta posição, que corresponde a uma taxa de seleção de 25%, afetando diretamente a eficiência seletiva e podendo levar a ganhos genéticos divergentes do esperado. Dessa forma, os genótipos selecionados pelo FAI-BLUP no Cenário I proporcionam ganhos genéticos desejados para todas as características consideradas nesse estudo.

A testemunha comercial, Tomate Fascínio, foi a melhor ranqueada pelo índice em ambos os cenários. Possivelmente, isso se deve ao fato de ser um híbrido simples já melhorado geneticamente para as características avaliadas. Os demais genótipos avaliados podem envolver híbridos triplos e duplos que possuem desempenho agronômico inferior quando comparado ao híbrido simples. Além disso, as linhagens envolvidas nos cruzamentos não são comerciais e podem ter as características consideradas no índice com natureza inferior quando comparadas com a testemunha.

4.5. CONCLUSÃO

Este estudo demonstrou como o FAI-BLUP pode ser afetado pela dependência espacial do campo experimental na seleção de genótipos superiores para múltiplas características em tomateiro. No caso do Cenário II, a dependência espacial levou a classificação errônea dos genótipos e reduziu a assertividade na seleção pelo FAI-BLUP. Já o Cenário I se mostrou o mais favorável para implementação do índice com posterior seleção de genótipos de tomateiro para múltiplas características. Assim, o estudo e seleção de modelos para cada característica é um passo importante para implementar e obter melhor desempenho do FAI-BLUP em direção ao ideótipo delineado para o tomateiro. Os genótipos selecionados podem avançar para as etapas seguintes do programa de melhoramento.

4.6. REFERÊNCIAS

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification”. IEEE TRANSACTIONS ON AUTOMATIC CONTROL 19 (6): 716–23. https://doi.org/https://doi.org/10.1109/TAC.1974.1100705.

Andrade, Mario Henrique Murad Leite, Claudio Carlos Fernandes Filho, Maiara Oliveira Fernandes, Abel Jamir Ribeiro Bastos, Marcio Lisboa Guedes, Tiago de Souza Marçal, Flavia Maria Avelar Gonçalves, Cesar Augusto Brasil Pereira Pinto, e Lincoln Zotarelli. 2020. “Accounting for spatial trends to increase the selection efficiency in potato breeding”. Crop Science 60 (5): 2354–72. https://doi.org/10.1002/csc2.20226.

Alvarenga MAR, Lima LA, Faquin V, Pereira GM (2013) Irrigação e fertirrigação. *In*: Alvarenga MAR (2013) Tomate: Produção em campo, casa de vegetação e hidroponia. 2. ed. Lavras: Editora Universitária de Lavras. p. 131-180.

Bartlett, M. S. Nearest Neighbour Models in the Analysis of Field Experiments. J. R. Statist. Soc. B, v. 40, n. 2, p. 147–174, 1978.

Bernardeli, Arthur, João Romero Amaral Santos de Carvalho Rocha, Aluízio Borém, Rodrigo Lorenzoni, Rafael Aguiar, Jéssica Nayara Basílio Silva, Rafael Delmond Bueno, et al. 2021. “Modeling spatial trends and enhancing genetic selection: An approach to soybean seed composition breeding”. *Crop Science* 61 (2): 976–88. https://doi.org/10.1002/csc2.20364.

Bojarian, Mohammad, Hossein Ali Asadi-Gharneh, e Maryam Golabadi. 2018. “Factor analysis, stepwise regression and path coefficient analyses of yield, yield-associated traits, and fruit quality in tomato”. *International Journal of Vegetable Science* 25 (6): 542–53. https://doi.org/10.1080/19315260.2018.1551260.

Botega, Gustavo Pucci. 2019. “EFICIÊNCIA DO ÍNDICE FAI-BLUP NA SELEÇÃO DE GENÓTIPOS BOURBON”. Dissertação de mestrado, Lavras: Universidade Federal de Lavras.

Burgueño, Juan. 2018. “Spatial Analysis of Field Experiments”. Em *Applied Statistics in Agricultural, Biological, and Environmental Sciences*, editado por Barry Glaz e Kathleen M Yeater, 319–44. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America. https://doi.org/10.2134/appliedstatistics.2016.0011.

Butler, D G, B R Cullis, A R Gilmour, B J Gogel, e R Thompson. 2023. “ASReml-R Reference Manual Version 4.2”. Hemel Hempstead, UK.: VSN International Ltd. https://asreml.kb.vsni.co.uk/article-categories/asreml-r-resources/.

Cavanaugh, Joseph E., e Andrew A. Neath. 2019. “The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements”. *WIREs Comput Stat.* e1460 (maio). https://doi.org/10.1002/wics.1460.

Copati, Mariane Gonçalves Ferreira, Françoise Dalprá Dariva, Felipe de Oliveira Dias, João Romero do Amaral Santos de Carvalho Rocha, Herika Paula Pessoa, Gabriella Queiroz de Almeida, Pedro Crescêncio Souza Carneiro, e Carlos Nick. 2021. “Spatial modeling increases accuracy of selection for Phytophthora infestans-resistant tomato genotypes”. *Crop Science* 61 (6): 3919–30. https://doi.org/10.1002/csc2.20584.

Copati, Mariane Gonçalves Ferreira, Herika Paula Pessoa, Françoise Dalprá Dariva, Manoel Nelson de Castro Filho, e Carlos Nick. 2024. “Tomato families possessing resistance to late blight also display high-quality fruit”. *Acta Scientiarum - Agronomy* 46 (1). https://doi.org/10.4025/actasciagron.v46i1.66790.

Costa, Cristina dos Santos Ribeiro, Maria Auxiliadora Coêlho de Lima, Francisco Pinheiro Lima Neto, Antonio Elton da Silva Costa, João Claudio Vilvert, Luiza Suely Semen Martins, e Rosimar dos Santos Musser. 2023. “Genetic parameters and selection of mango genotypes using the FAI-BLUP multitrait index”. *Scientia Horticulturae* 317 (julho). https://doi.org/10.1016/j.scienta.2023.112049.

Cullis, Brian R., A. B. Smith, e N. E. Coombes. 2006. “On the design of early generation variety trials with correlated data”. *Journal of Agricultural, Biological, and Environmental Statistics* 11 (4): 381–93. https://doi.org/10.1198/108571106X154443.

Eshed, Y., and D. Zamir. 1995. “An Introgression Line Population of Lycopersicon Pennellii in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield- Associated QTL.” Genetics 141 (3). Genetics Society of America: 1147–1162. https://doi.org/10.1093/genetics/141.3.1147.

Filgueira FAR, Obeid PC, Morais HJ, Santos WV, Fontes RR (1999) Tomate tutorado. In: Ribeiro AC, Guimarães PTG, Alvarez VHV Recomendações para o uso de corretivos e fertilizantes em Minas Gerais - 5ª Aproximação. Viçosa, Editora SBCS. p. 187-188.

Furlani PR, Faquin V, Alvarenga MAR, Seno S (2013) S. Produção em substratos e em hidroponia. In: Alvarenga MAR (2013) Tomate: Produção em campo, casa de vegetação e hidroponia. 2. ed. Lavras: Editora Universitária de Lavras. p. 245-273

Gardner, RG, and Panthee, D.R. 2010a. NC 1 CELBR and NC 2 CELBR: Early Blight and Late Blight-resistant Fresh Market Tomato Breeding Lines. Hort Science 45(6): 975-976.

Gardner, R.G., and Panthee, D.R. 2010b. “Plum Regal” fresh-market plum tomato hybrid and its parents, NC 25p and NC 30p. HortScience 45:824–825.

Gilmour, Arthur R, Brian R Cullis, e Arūnas P Verbyla. 1997. “Accounting for natural and extraneous variation in the analysis of field experiments”. *Source: Journal of Agricultural, Biological, and Environmental Statistics* 2 (3): 269–93. http://www.jstor.org/stable/1400446http://www.jstor.org/page/info/about/policies/terms.jsp.

Henderson, C R. 1975. “Best linear unbiased estimation and prediction under a selection model”. *Biometrics* 31 (2): 423–47.

Hoefler, Raegan, Pablo González-Barrios, Madhav Bhatta, Jose A.R. Nunes, Ines Berro, Rafael S. Nalin, Alejandra Borges, et al. 2020. “Do Spatial Designs Outperform Classic Experimental Designs?” *Journal of Agricultural, Biological, and Environmental Statistics* 25 (4): 523–52. https://doi.org/10.1007/s13253-020-00406-2.

Kaiser, Henry F. 1958. “The varimax criterion for analytic rotation in factor analysis”. *Psychometrika* 23 (3): 187–200. https://doi.org/https://doi.org/10.1007/BF02289233.

Meier, Carine, Volmir Sergio Marchioro, Daniela Meira, Tiago Olivoto, e Luís Antônio Klein. 2021. “Genetic parameters and multiple-trait selection in wheat genotypes”. *Pesquisa Agropecuaria Tropical* 51. https://doi.org/10.1590/1983-40632021v5167996.

Neath, Andrew A., e Joseph, E. Cavanaugh. 2012. “The Bayesian information criterion: Background, derivation, and applications”. *WIREs Comput Stat* 4 (2): 199–203. https://doi.org/10.1002/wics.199.

Nick, Carlos, e Silva, Derly José Henriques. 2016. Melhoramento de tomate. In: NICK, C.; BOREM, A. Melhoramento de hortaliças. Viçosa: UFV cap.13, p.396-431.

Oliveira, Rebeca Lourenço de, Ronaldo Silva Gomes, Cleverson Freitas de Almeida, Ronaldo Machado Júnior, João Romero A.S.de C. Rocha, Derly José Henriques da Silva, e Pedro Crescêncio Souza Carneiro. 2021. “Multitrait selection of pumpkin genotypes aimed at reducing the growth habit and improving seed production”. *Crop Science* 61 (3): 1620–29. https://doi.org/10.1002/csc2.20386.

Olivoto, Tiago, e Alessandro Dal'Col Lúcio. 2020. "metan: An R package for multi-environment trial analysis". *Methods in Ecology and Evolution* 11 (6): 783–89. https://doi.org/10.1111/2041-210X.13384.

Patterson, H D, e R Thompson. 1971. "Recovery of inter-block information when block sizes are unequal". *Biometrika* 58 (3): 545–54.

Piepho, H. P., J. Möhring, A. E. Melchinger, e A. Büchse. 2008. "BLUP for phenotypic selection in plant breeding and variety testing". *Euphytica* 161 (1–2): 209–28. https://doi.org/10.1007/s10681-007-9449-8.

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Resende, Marcos Deon Vilela, e José Alfredo Sturion. 2001. "Análise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de séries temporais empregando REML/BLUP ao nível individual". Colombo, PR. www.cnpf.embrapa.br.

Rocha, João Romero do Amaral Santos de Carvalho, Juarez Campolina Machado, e Pedro Crescêncio Souza Carneiro. 2018. "Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy". *GCB Bioenergy* 10 (1): 52–60. https://doi.org/10.1111/gcbb.12443.

Salvador, Felipe Vicentino, Gabriela dos Santos Pereira, Michel Henriques de Souza, Laiza Maria Bendia da Silva, Alice Silva Santana, Igor Gonçalves de Paula, Skarlet de Marco Steckling, et al. 2022. "Correcting experimental data for spatial trends in a common bean breeding program". Crop Science 62 (2): 825–38. https://doi.org/10.1002/csc2.20703.

Schwarz, Gideon. 1978. "Estimating the dimension of a model". *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Silva, Caique Machado, Victor Silva Signorini, Saulo Fabrício da Silva Chaves, Diana Jhulia Palheta de Souza, Gabriel Wolter Lima, Cleiton Renato Casagrande, Henrique Caletti Mezzomo, João Paulo Oliveira Ribeiro, e Maicon Nardino. 2024. "Modeling spatial trends and selecting tropical wheat genotypes in multi-environment trials". Crop Breeding and Applied Biotechnology 24 (2): 47582421. https://doi.org/10.1590/1984.

Velazco, Julio G., María Xosé Rodríguez-Álvarez, Martin P. Boer, David R. Jordan, Paul H.C. Eilers, Marcos Malosetti, e Fred A. van Eeuwijk. 2017. "Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model". Theoretical and Applied Genetics 130 (7): 1375–92. https://doi.org/10.1007/s00122-017-2894-4.

5. CAPÍTULO 2:

REMOTE SENSING AND MACHINE LEARNING TECHNIQUES FOR HIGH THROUGHPUT PHENOTYPING OF LATE BLIGHT-RESISTANT TOMATO PLANTS IN OPEN FIELD TRIALS

Felipe de Oliveira Dias [a]*, Domingos Sarvio Magalhães Valente[b], Carolina Tavares de Oliveira[b], Françoise Dalprá Dariva[a], Mariane Gonçalves Ferreira Copati [a] and Carlos Nick[a]

[a] *Departamento de Agronomia, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Campus Universitário, Viçosa, Minas Gerais, Brazil, 36570-900.*

[b] *Departamento de Engenharia Agrícola, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Campus Universitário, Viçosa, Minas Gerais, Brazil, 36570-900.*

* Corresponding author: felipedeoliveiradias@gmail.com

Highlights

1. A multispectral camera was used to phenotype tomato lines in the field;
2. A Random Forest model was created to predict late blight severity in tomato plots;
3. The model differentiated resistant from susceptible tomato lines;
4. The methodology enables high-throughput phenotyping in the field.

Abstract

The selection of late blight resistant genotypes of tomato requires many evaluations on the field. The use of machine learning models to assess late blight severity based on images from multispectral cameras onboard of unmanned aerial vehicles (UAVs) can bring efficiency and quickly in evaluations. In this work, we use remote sensing and machine learning techniques to assess late blight resistance of tomato lines grown in open field conditions. Seventy-six tomato lines, including two resistant and one susceptible line, were used to quantify late blight severity. Plants were arranged according to a replicated check design in a total of 132 experimental plots. Tomato plants were artificially inoculated with a *Phytophthora infestans* zoospore suspension. Multispectral images were obtained using an unmanned aerial vehicle. We calculated vegetation indexes (VI) using the images, which were the basis for building the Random Forest models used to predict disease severity. Two methodologies were used to predict late blight severity: Methodology 1, which used only the images from the last day of evaluation, and Methodology 2, which used the images from four days of evaluation. For Methodology 1 and 2, determination coefficients of 0.81 and 0.93 were obtained for the test set, respectively. Methodology 2 was used to predict late blight severity of the 132 field plots. Tomato plots were sorted from lowest to highest predicted severity. Resistant plots were ranked first indicating consistency of prediction. We therefore recommend Methodology 2 as a fast and practical way to predicted late blight severity in breeding populations.

Keywords: *Solanum lycopersicum*; *Phytophthora infestans*; tomato breeding, UAV images, Random Forest model

5.1. INTRODUCTION

The main challenge of growing tomatoes in tropical regions is to have to deal with the high number of pests and diseases affecting tomato plants throughout the entire growing cycle. Bergougnoux (2014) reported the likely existence of more than 200 pests and diseases causing damage to tomato plants, which results in the adoption of immediate control measures, such as the application of pesticides, as a way to minimize crop loss. Among the diseases that affect tomato crops, late blight, caused by the oomycete *Phytophthora infestans* Mont de Bary, has shown high damage

potential. This disease affects several plant organs, and in conditions of inadequate management, it rapidly compromises the whole aerial part of the plant (Duarte, Zambolim and Jesus Junior 2007; Nowicki, Kozik and Foolad 2013). Disease progression can lead to plant death and hence 100% production loss under favorable environmental conditions.

Late blight management at field conditions involves preventive fungicide applications, which substantially increases production costs. According to Zanotta et al. (2016), fungicide applications for late blight control account for 15 to 20% of a tomato field's total production cost. A more sustainable approach to control late blight is the use of resistant plant materials (Nowicki, Kozik and Foolad 2013; Park et al. 2005). However, in tropical conditions, the number of cultivars displaying resistance to *P. infestans* isolates is worryingly low. Such restricted number of cultivars is explained by the difficulty of phenotyping large tomato populations for late blight resistance, with workforce costs being the primary limiting factor (Chawade et al. 2019). Many personnel is required for more accurate phenotyping, making the process laborious, costly, and time-consuming. Moreover, it requires experienced personnel to the precise identification of disease symptoms and pathogen signs, leading to a substantial subjectivity of the process (Chawade et al. 2019; Cruz et al. 2019), which in turn could result in inaccurate disease severity scores. Besides, there is a need to limit the number of genotypes tested and reduce the number of replicates to a minimum, especially during initial breeding stages.

In order to overcome possible mistakes that are common to the existing phenotyping methodologies, Corrêa, Bueno Filho and Carmo (2009) developed a disease severity scale to assess late blight progress rate in tomato plants. Yet, its use in breeding programs is limited due to the large number of genotypes assessed during initial breeding stages. Another limitation related to the use of this disease severity scale is the accuracy of disease severity scores based on visual observations.

Other phenotyping techniques available to tomato breeding trials involve the use of images obtained from satellites, UAVs, or ground-based vehicles (Chawade et al. 2019). Works using spectral radiance of tomato plants to identify diseases are available in the literature. Zhang, Liu and O'Neill (2002) discriminated tomato plants displaying late blight symptoms from healthy ones in advanced stages of disease infection through reflectance data obtained by a spectroradiometer-GER2600. Xie et al. (2015) were able to differentiate two diseases affecting tomato plants using

hyperspectral images of tomato leaves. Images were taken by a device equipped with a camera (C8484-05, Hamamatsu, Japan), an imaging spectrograph (V10E-QE, Specim, Finland), a lens (OLE-23), and two light sources (150 W- tungsten halogen lamps).

Modern drones and sensors have facilitated the acquisition of multispectral images for several purposes. According to Deng et al. (2018) it is necessary to study the ability of sensors in determining phenotypes of individuals because spectral responses may vary depending on the sensor. Machine learning algorithms can be used to analyze spectral imaging data, however, little is known about the quantification and prediction of plant diseases through this technique (Singh et al. 2021), which makes its adoption difficult in breeding initiatives.

Studies for efficient and precise diagnosis of disease damage in crops are currently in development. Prasad et al. (2022) suggested a two-step methodology. First, low-resolution images from apple plants were used to train an identifier to locate potentially diseased regions inside the images. Then, high-resolution images were used to train a classifier for accurate plant health diagnosis. To validate this methodology, a DJI Marvic 2 drone, equipped with a 1" CMOS sensor, flew over an apple orchard to take low-resolution pictures. The identifier located the diseased plant parts and the classifier reported the health status of the plants through high-resolution images. According to these researchers, the methodology was highly accurate in separating sick from healthy plants.

Automated phenotyping of tomato plants for resistance to *P. infestans* could allow the assessment of a large number of plant materials in a short period (Luvisi, Ampatzidis and De Bellis 2016). This approach would also reduce the subjectivity of disease assessments, increasing genotyping selection accuracy (Chawade et al. 2019). With the goal of enabling high-throughput phenotyping for *P. infestans* resistance in tomato plants, this study verified if late blight severity scores could be accurately predicted by machine learning models that use image data from multispectral cameras onboard UAVs. We used remote sensing images, acquired with a MicaSense RedEdge-MX multispectral camera onboard of a DJI Matrice 100 drone, and the Random Forest machine learning algorithm, to estimate late blight severity in tomato plants grown in the Viçosa field 2019.

5.2. MATERIAL AND METHODS

5.2.1. Plant material

We tested 71 introgression lines (ILs) derived from a cross between the processing tomato cultivar M82 (*Solanum lycopersicum*) and the wild tomato accession LA716 (*Solanum pennellii*) (Eshed and Zamir 1995). Five other tomato materials, used as checks, were also included in the trial: the cultivar Santa Clara, considered susceptible to *P. infestans*; the line NC 25P carrying the *Ph-3* resistance gene (Gardner and Panthee 2010a); and the line NC 1 CELBR carrying *Ph-2* and *Ph-3* genes, both conferring resistance to *P. infestans* isolates (Gardner and Panthee 2010b); the processing tomato variety M82 and the *S. pennellii* accession LA716.

5.2.2. Site and field conditions

The experiment was carried out in 2019 in the Research and Extension Farm Unit *Horta Velha* belonging to the Department of Agriculture at Universidade Federal de Viçosa in the municipality of Viçosa, Minas Gerais State, Brazil, located at 20.75° S and 42.88° W, 648.74 m of altitude. According to Köppen's classification, regional climate is of type Cwb, mesothermic, with rainy summers and dry winters.

Tomato seedlings were grown conventionally and fertilized according to Furlani et al. (2013). Soil preparation followed conventional farming procedures required for the tomato crop. Fertilization was based on soil fertility results and recommendations of Filgueira et al. (1999). Fertilizer application was partitioned according to Alvarenga (2013). Standard production practices for the tomato crop were carried out as needed. Table 1 contains a calendar with tomato phenological phases described by Shamshiri et al. (2018).

Table 13: Phenological phases in tomato. Adapted from Shamshiri et al. (2018).

Tomato Developmental stage	Duration* (days)
Early growth	25-30
Vegetative	20-25
Flowering	20-30
Fruit formation	20-30
Mature fruiting	15-20

*Depends on the tomato cultivar.

5.2.3. Experimental design

The experiment was arranged according to a replicated checks design (Cruz, Carneiro, and Regazzi 2014). Experimental plots consisted of five tomato plants in a row, with only the central three being used to quantify late blight severity. An experimental field scheme is shown in Figure 1.

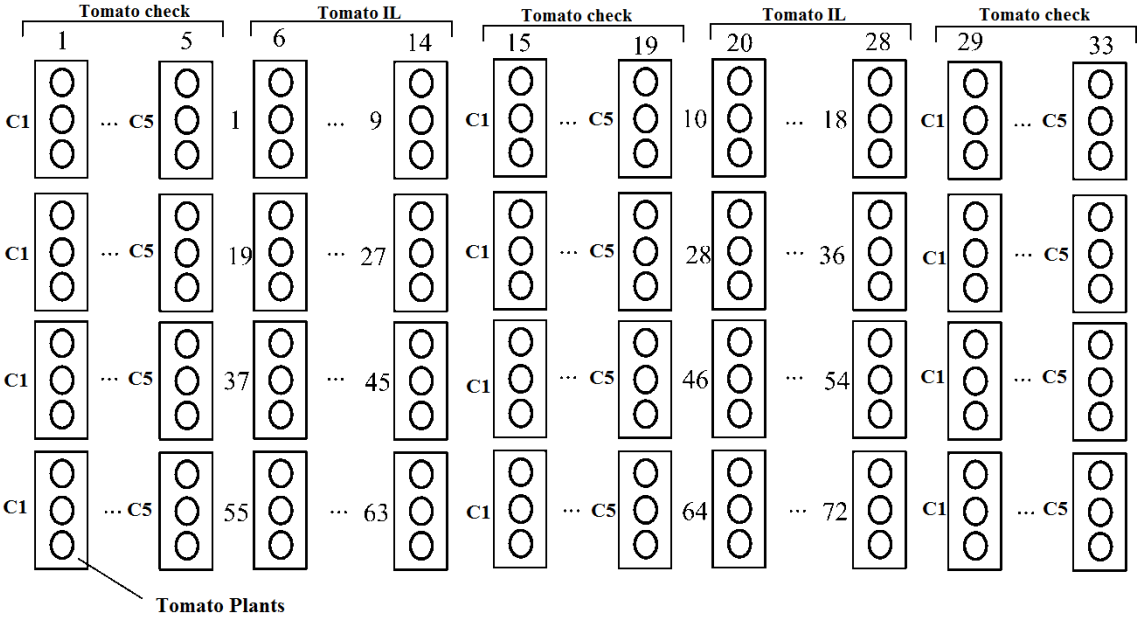

Figure 1: Experimental field scheme showing plot arrangement. Numbers 1 to 72 represent the tomato ILs; C1 to C5 represent the tomato checks used. Plants were assessed according to their resistance to *P. infestans*.

Tomato ILs were numerically identified from 1 to 72 and were not replicated. Checks were identified from C1 to C5 and were replicated 12 times. The experiment consisted of 33 rows containing 132 plots in total. Plot number 60 was formed by a mixture of genotypes. Between-row and in-row spacing was 1.00 x 0.60 m, respectively, so that the total area of a single plot was 3 m².

5.2.4. Isolate preparation and application

All genotypes were artificially inoculated with a *P. infestans* zoospore solution. We used the same methodology described by Abreu et al. (2008) with few modifications. *P. infestans* isolates were collected in commercial tomato fields of Coimbra - MG, Brazil. The collected *P. infestans* isolates were mixed up to form the inoculum solution used to infect the plants 45 days after field transplanting. Isolate concentration was adjusted to 1 x 10^5 sporangia per milliliter of suspension using a

hemacytometer. Equally spaced overhead sprinklers, placed throughout the whole experimental field, were turned on every day in the afternoon period to promote leaf watering, an adequate condition for late blight development. Chemical applications were suspended seven days before inoculation.

After inoculation, trained personnel (one for each block) attributed disease severity scores for each leaf on each plant according to the diagrammatic scale proposed by Corrêa, Bueno Filho and Carmo (2009) (Figure 2).

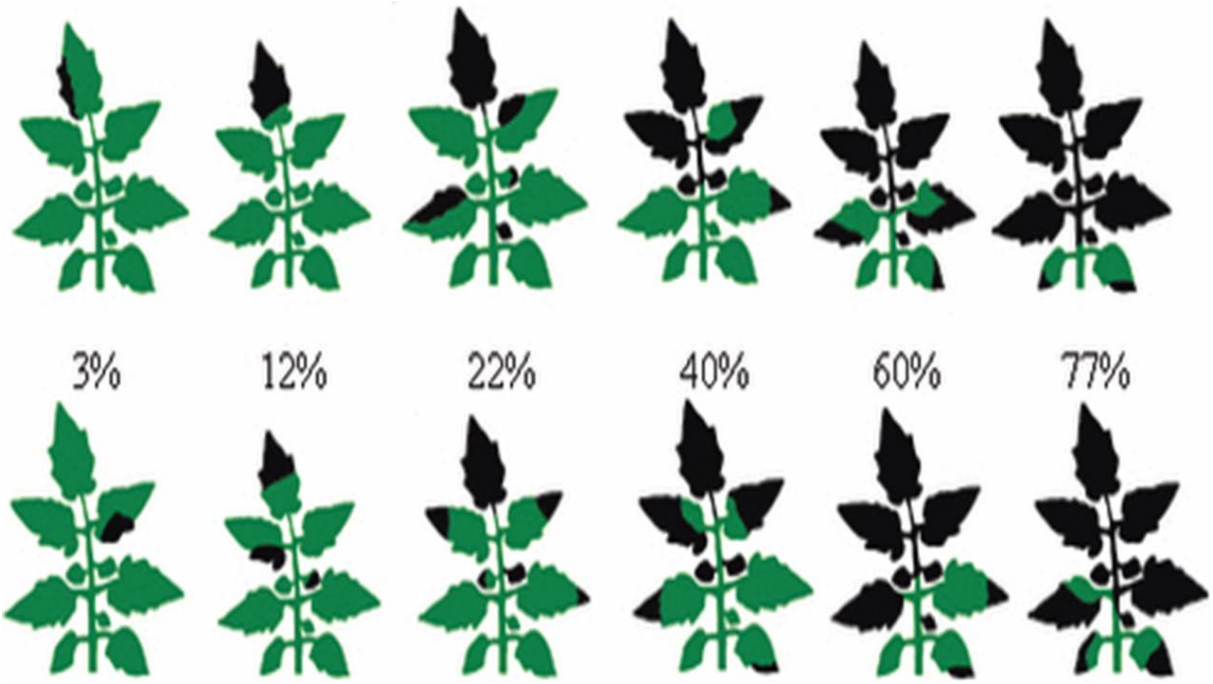

Figure 2: Diagrammatic scale used to assess late blight severity observed in tomato plants. Late blight lesions are highlighted in black. Evaluations may be subjective. Source: Corrêa, Bueno Filho and Carmo (2009).

Late-blight severity scores were recorded every three days for 18 days. Late blight severity ratings were then used to estimate the area under the disease progress curve (AUDPC) as according to (Campbell and Madden 1990).

5.2.5. Image data collection

Image data were acquired with a MicaSense RedEdge-MX multispectral camera, focal length of 5.4 mm, and sensor size of 4.8 x 3.6 mm (MicaSense 2021), on a UAV (DJI Matrice 100, 2021). The camera has five complementary metal-oxide-semiconductor sensors providing information about the blue (475 nm), green (560 nm), red (668 nm), *RedEdge* (717 nm), and near infra-red (840 nm) spectral bands. Drone-

based images were taken on June 28th, July 1st, July 4th, and July 8th, between 11 am and 1 pm. Each day, a single image taken from a flying altitude of 50 m, with the drone motionless in the sky, was sufficient to capture the whole experimental field (33 x 12 m) with a spatial resolution of 5 cm per pixel. A Micasense camera contains an integrated GNSS sensor and a DSL-2 module that quantifies irradiance, sun-ray angles. Radiometric calibration of the equipment was performed according to the manufacturer's instructions through the camera's software.

Image processing consisted of creating contour polygons for each tomato plot using the QGIS software, version 2.18 (QGIS Development Team 2019), according to the scheme shown in Figure 1. A total of 132 polygons (cropped images) were created per image, with each polygon representing a single plot (3 x 1 m). After QGIS processing, each polygon was transformed into an independent cropped image. A cropped image was formed by the tomato plants, soil, and shadow. For each cropped image, we obtained the corresponding blue (Blue band - B), green (Green band - G), red (Red band - R), *RedEdge* (RedEdge band - RE), and infrared (Near-Infrared band - IR) bands, as shown in Figure 3.

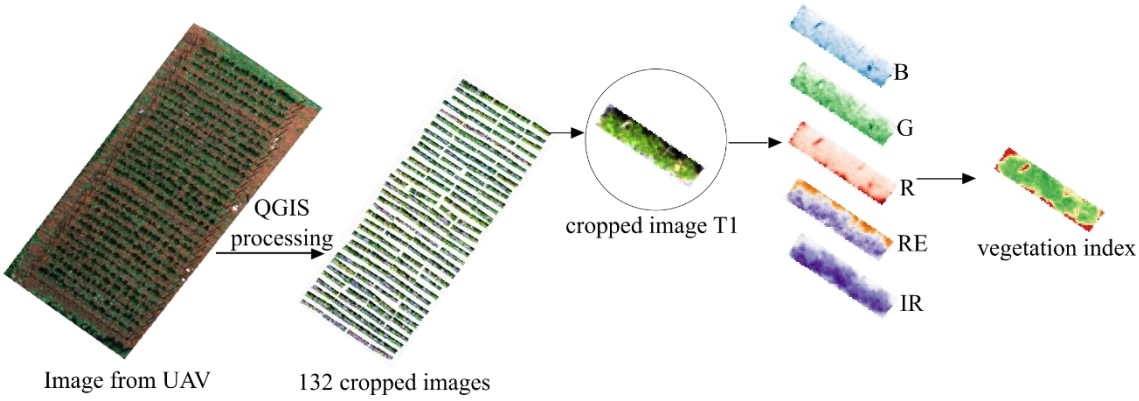

Figure 3: Steps of image processing to calculate vegetation indexes of each treatment. B: blue band, G: green band, R: red band, RE: red edge band, IR: infrared band.

Based on such spectral bands, we calculated the vegetation indexes presented in Table 2.

Table 2 – Vegetation indexes (VI) estimated for each treatment.

VI	Calculation method	Source
NDVI	$\mathrm{NDVI} = (IR - \mathrm{R})/(\mathrm{IR} + \mathrm{R})$	Rouse et al. (1974)
GNDVI	$\mathrm{GNDVI} = (\mathrm{IR} - \mathrm{G})/(\mathrm{IR} + \mathrm{G})$	Gitelson, Kaufman and Merzlyak (1996)
SAVI	$\mathrm{SAVI} = 1.5(\mathrm{IR} - \mathrm{R})/(0.5 + \mathrm{IR} + \mathrm{R})$	Rondeaux, Steven and Baret (1996)
NDVIRE	$\mathrm{NDVIRE} = (\mathrm{IR} - \mathrm{RE})/(\mathrm{IR} + \mathrm{RE})$	Gitelson and Merzlyak (1997)
MCARI1	$\mathrm{MCARI1} = 1.2[2.5(\mathrm{IR} - \mathrm{R}) - 1.3(\mathrm{IR} - \mathrm{G})]$	Daughtry et al. (2000)
MTVI1	MTVI1=1.2[1.2(IR − G) – 2.5(R − G)]	Haboudane et al. (2004)
SR	$\mathrm{SR} = \mathrm{IR}/\mathrm{R}$	Jordan (1969)
NGRDI	$\mathrm{NGRDI} = (\mathrm{G} - \mathrm{R})/(\mathrm{G} + \mathrm{R})$	Tucker (1979)
RVI	$\mathrm{RVI} = \mathrm{R}/\mathrm{IR}$	Jordan (1969)
NRVI	$\mathrm{NRVI} = (\mathrm{RVI} - 1)/(\mathrm{RVI} + 1)$	Baret and Guyot (1991)

As the cropped images also carried soil information, they were filtered using the Normalized Difference Vegetation Index (NDVI). With the NDVI values, we created filtering masks for each cropped image according to the following rule: NDVI > 0.20, NDVI > 0.30, NDVI > 0.40, NDVI > 0.50, NDVI > 0.60. This mask was used to filter the original image for each treatment. Figure 4 shows an example of a cropped image modified by the filtering masks. Black pixels mean that the pixel value is equal to 1 (met the rule). White pixels mean that the pixel value is equal to 0 (did not meet the rule). Considering this pixel criterion, we then created one image for each NDVI mask. Vegetation indexes were calculated based on filtered images.

Figure 4: Example of masks created for a single cropped image based on the NDVI index. NDVI values greater than the established threshold mean that pixel value is 1 (black). Otherwise, the pixel value is 0 (white).

Mean values of all vegetation indexes, including the NDVI index, were estimated based on the filtering masks. To calculate the average vegetation indexes, we considered only the pixels valuing 1 on the referred mask. We used this strategy in order to remove irrelevant pixels or noise (soil and shadow) on each cropped image. Vegetation index averages for each image (treatment) were estimated based on NDVI indexes for each filtered image. Each dataset feature was formed by vegetation index averages for each NDVI filter. Therefore, for each mask, we obtained a dataset consisting of 132 rows (cropped images) and 10 columns (means of each vegetation index), totaling five datasets for each day. For each cropped image, we also scored late blight severity based on the diagrammatic scale from the last day of evaluation (July 8th). The Python 3.7 language was used here to process the vegetation indexes and generate the datasets.

5.2.6. Machine learning modeling

For modeling, we used the Random Forest algorithm for regression in *Python* 3.7, and *Scikit Learn* library to estimate predict severity (Pedregosa et al. 2011). Two different methods were assessed here. In the first method, we considered as features only the vegetation index data (independent variables) from July 8th to predict late blight severity on this same day. In such approach, five datasets corresponding to the five NDVI masks applied for data filtering were used. In the second method, we considered as features the vegetation index data (independent variables) from June 28th, July 1st, July 4th, and July 8th obtained for each mask. In this case, for each mask, a new dataset containing 132 rows (cropped images) and 40 columns (average vegetation index for each day on each mask) was formed. On both methods, we intended to create models with the best results for disease severity prediction so that on each created model, seven different steps (A, B, C, D, E, F, and G) were taken as described in Figure 5.

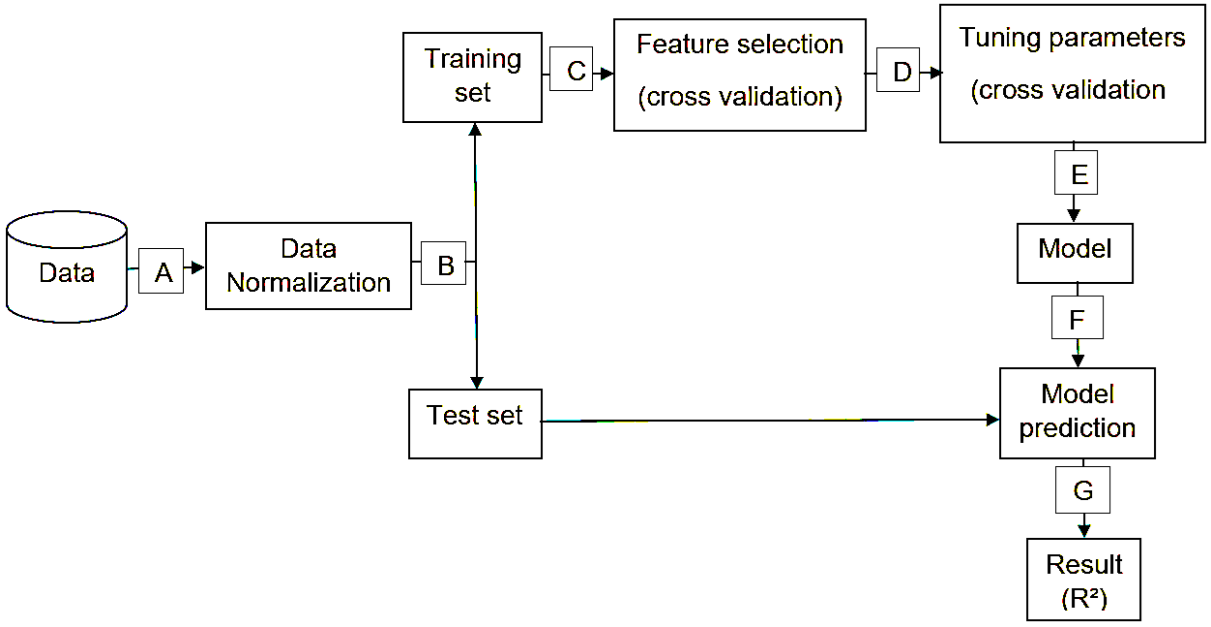

Figure 5: Flowchart of the methodology used to build and validate the machine learning models.

In Figure 5, in step A, data were normalized to range from 0 to 1 (minimum and maximum). This step was taken because the minimum-to-maximum interval between variables was different. Next, in step B, we randomly divided the dataset into a training set and a test set. The training set corresponded to 80% of the original dataset, while the test set corresponded to 20%. In step C, we intended to select features from the training set. Feature selection is a process at which we automatically select the independent variables that most contribute to the output variable (Brownlee 2019). The backward elimination algorithm was adopted for feature selection, with the importance of features being used as the criteria. The criteria used to stop feature elimination was the optimum determination coefficient value in cross-validation with five k-folds (R^2-CV) in the training set. In step D, we optimized the hyperparameters in order to improve model performance. The hyperparameters (parameters defined by the user), optimized in the Random Forest model, were: number of trees (number of estimators), maximum depth of each tree, the minimum number of samples on each node (MSL), the minimum number of samples needed to split an internal node (MSS). The optimization of hyperparameters was based on determination coefficient results generated with five k-folds- cross-validation (R^2-CV) in the training set. After hyperparameter optimization, in step E, a model containing all training set (80 % of the original dataset) was created for each mask. In this way, we calculate the determination coefficient (R^2-CV) and the root mean square error (root mean square error of cross-validation - RMSE-CV) of the

cross-validation. After obtaining the final models using the training set, we applied each model on the test set to calculate the determination coefficients of the test set (R^2) and the root mean square error (RMSE).

We selected the fittest model from both methodologies based on determination coefficient values in cross-validation (R^2-CV) and RMSE-CV values of filtering masks as well as R^2 and RMSE of each model. The fittest model was then used to predict late blight severity for each of the 132 polygons (experimental plots). Polygons were then ranked based on their disease-predicted severity. We considered resistant those lines with lowest predicted severity scores.

Overall, this study aimed at using droned-based images of a tomato field to predict the area under the disease progress curve for late blight in different tomato lines. Predictions were based on the best machine-learning model using different vegetation indexes.

5.3. RESULTS

In Figure 6 is shown the behavior of average late-blight severity scores on each day of evaluation.

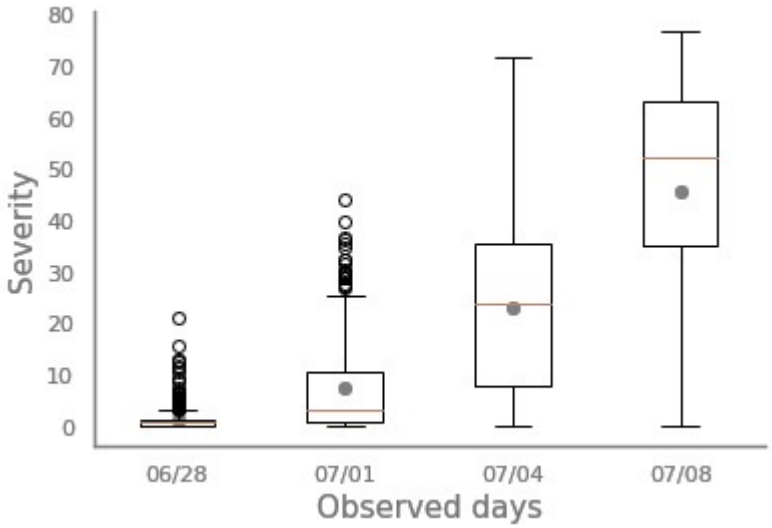

Figure 6: Average severity scores recorded in tomato genotypes assessed for late-blight resistance at field conditions.

It can be verified that on June 28th (first day of evaluation), just a few disease severity scores were superior to 20%. On this day, the average severity score was only 1.4%. We also observed a tendency of increase in disease severity means overtime. On the last day of evaluation, the average disease severity score was 45.56%.

After *P. infestans* inoculation, we verified a decrease in plant vigor overtime. Such response is clearly demonstrated by the descriptive statistical analyses presented in Table 3, which contains means and standard deviation values of all vegetation indexes for the NDVI > 0.20 mask. There was a tendency of decrease in the means overtime for most vegetation indexes.

Table 3 – Mean and standard deviation (SD) of vegetation indexes (VI) overtime considering the NDVI > 0.20 mask.

VI	June 28th		July 1st		July 4th		July 8th	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NDVI	0.722	0.069	0.719	0.053	0.678	0.055	0.683	0.052
NDVIRE	0.325	0.035	0.339	0.031	0.313	0.032	0.291	0.033
SAVI	1.083	0.103	1.079	0.079	1.018	0.082	1.025	0.078
GNDVI	0.649	0.042	0.686	0.034	0.638	0.035	0.641	0.034
MCARI1	14479	2408	25224	4405	12963	2804	13985	2736
MTVI1	14479	2408	25224	4405	12963	2804	13985	2736
NGRDI	0.179	0.065	0.090	0.048	0.091	0.048	0.090	0.046
RVI	0.170	0.051	0.168	0.038	0.198	0.042	0.193	0.038
NRVI	-0.762	0.069	-0.719	0.053	-0.678	0.055	-0.683	0.052
SR	8.010	1.770	7.177	1.306	6.026	1.121	5.970	1.301

A darker hue indicates a higher average. The lighter hue indicates lower average.

Table 4 shows Pearson's correlation coefficients between the average vegetation indexes for the NDVI > 0.20 mask and the disease severity scores on the last day of evaluation (July 8th).

Table 4 – Pearson's correlation coefficients between late-blight severity scores on the last day of evaluation (July 8th) and the vegetation indexes (VI) for the NDVI> 0.2 mask.

VI	June 28th	July 1st	July 4th	July 8th
NDVI	0.204	0.034	-0.275	-0.597
NDVIRE	0.237	0.110	-0.223	-0.624
SAVI	0.204	0.034	-0.275	-0.597
GNDVI	0.268	0.155	-0.223	-0.598
MCARI1	0.086	-0.144	-0.516	-0.800
MTVI1	0.086	-0.144	-0.516	-0.800
NGRDI	0.132	-0.142	-0.393	-0.638
RVI	-0.200	-0.040	0.237	0.550
NRVI	-0.204	-0.034	0.275	0.597
SR	0.224	-0.005	-0.455	-0.762

A darker hue indicates positive correlation. A lighter hue indicates a negative correlation.

Pearson's correlation coefficients were numerically higher on the last day of evaluation compared to the others. Moreover, there was a tendency for most vegetation indexes to correlate negatively with late-blight severity. The highest correlation coefficient values were observed for the Modified Chlorophyll Absorption in Reflectance Index 1 (MCARI1) and Modified Triangular Vegetation Index 1 (MTVI1). On Table 3, it is possible to verify that both indexes were very similar. It means that MCARI1 and MTVI1 indexes are strongly correlated, and the MTVI1 index was removed during the feature selection phase. Only two vegetation indexes on the last day of evaluation (July 8th) showed a positive correlation with disease severity scores (the Ratio Vegetation Index (RVI) and the NRVI). According to the equations shown in Table 2, these indexes have an inverse relationship with the infrared band (IR). The infrared band decreases with the disease progress, leading to an increase in the RVI and NRVI indexes, as shown in Table 3.

5.3.1. Method 1: AUDPC predictions using a single image

For each mask, we created a combination of features that maximized the models' performance. This feature selection was performed using the training set with cross-validation (5 k-folds) for each mask. The idea was to optimize determination coefficients of cross-validation (R^2-CV). Results are shown in Figure 7.

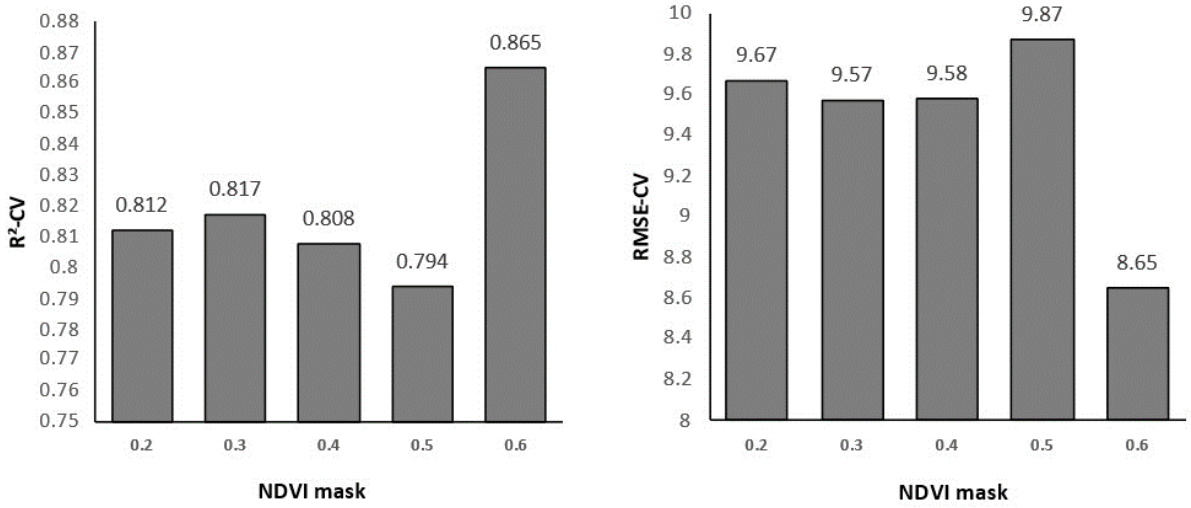

Figure 7: Average of the determination coefficients (R^2-CV) and root mean square error of cross-validation (RMSE-CV) during modeling, based on vegetation indexes on each NDVI mask. Here we used only the aerial image taken on the last day of evaluation (July 8th).

The highest average R^2-CV was 0.87 found for the NDVI > 0.6 mask. This mask presented the lowest average for RMSE-CV (8.65%). With the five models created (one for each mask), it was possible to predict late-blight severity scores using the test set (results are shown in Figure 8).

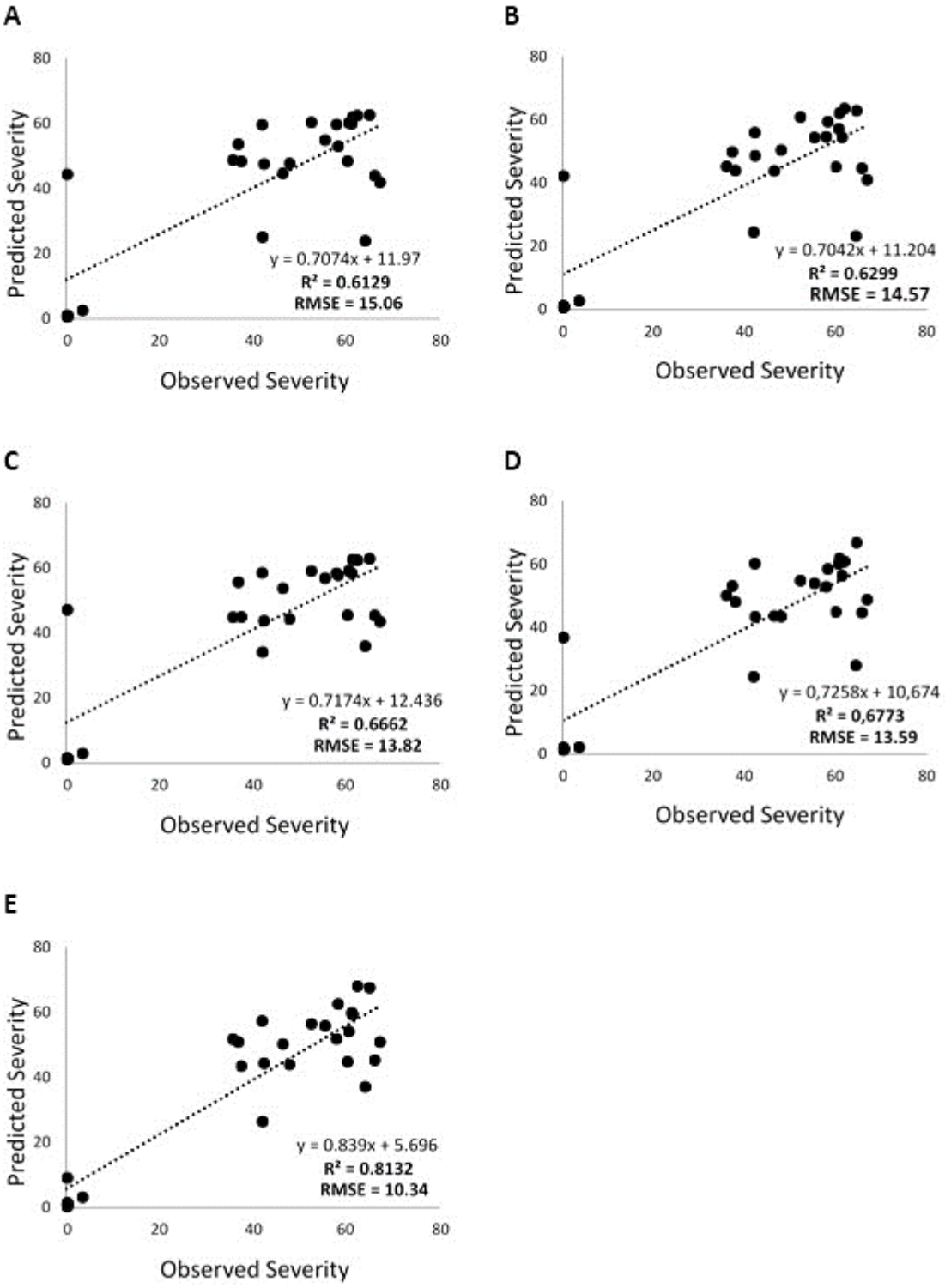

Figure 8: Prediction of late blight severity scores for the test set (data that did not participate in modeling). Models were based on vegetation indexes created using the information of a single aerial image taken on the last day of evaluation and the (A) NDVI > 0.2, (B) NDVI > 0.3, (C) NDVI > 0.4, (D) NDVI > 0.5 and (E) NDVI > 0.6 masks.

Based on the average determination coefficients of cross-validation, the best mask was NDVI > 0.60 (Figure 7). This mask created a model with an average R^2-CV of 0.87. With this model, we obtained a determination coefficient (R^2) for the test set of

0.81 and RMSE of 10.34%, according to Figure 8E.

Figure 9A shows in detail a multispectral image taken in the experimental field on July 08th. Santa Clara and NC1 CELBR check plots are shown in Figure 9B and Figure 9C, respectively.

Figure 9: Multispectral image on July 08th. Pixel size = 5 cm. **A** – experimental field; **B** –Santa Clara plot, susceptible check; **C** – NC1 CELBR plot, resistant check.

5.3.2. Method 2: AUDPC prediction using four images

Method 2 was created using vegetation index data from four days of evaluation, and it was used to predict disease severity scores on the last day of evaluation. The highest R^2-CV and the lowest RMSE-CV was 0.86 and 7.62%, respectively, for the NDVI>0.6 mask (Figure 10).

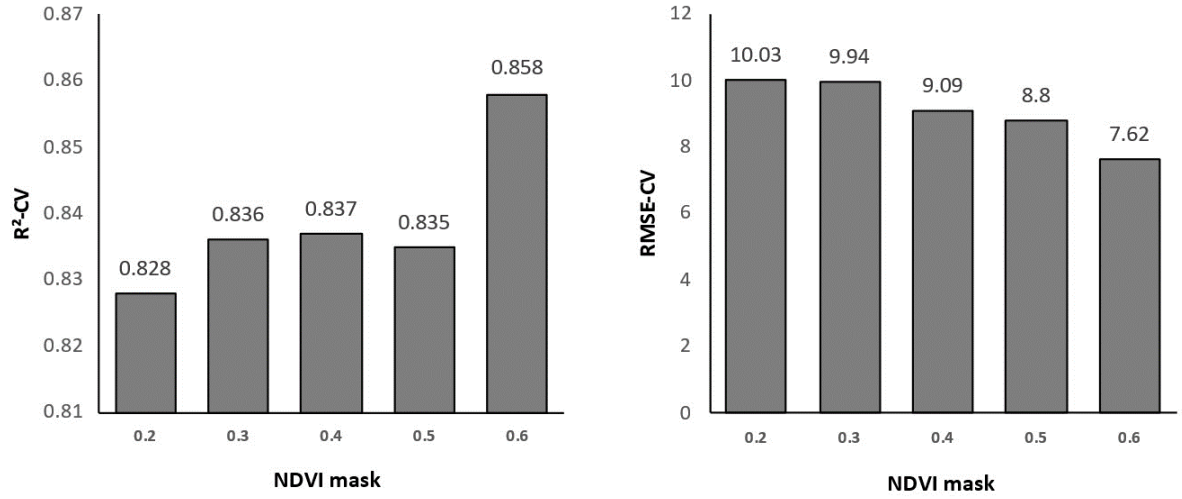

Figure 10: Average of the determination coefficients (R^2-CV) and the root mean square error of cross-validation (RMSE-CV) during modeling, based on vegetation indexes on each NDVI mask. We used information of four aerial images taken during the disease severity evaluation period to calculate the vegetation indexes here.

We predicted disease severity scores for the test set, and then we created the graphs (predicted severity versus observed severity) for each NDVI mask, as shown in Figure 11.

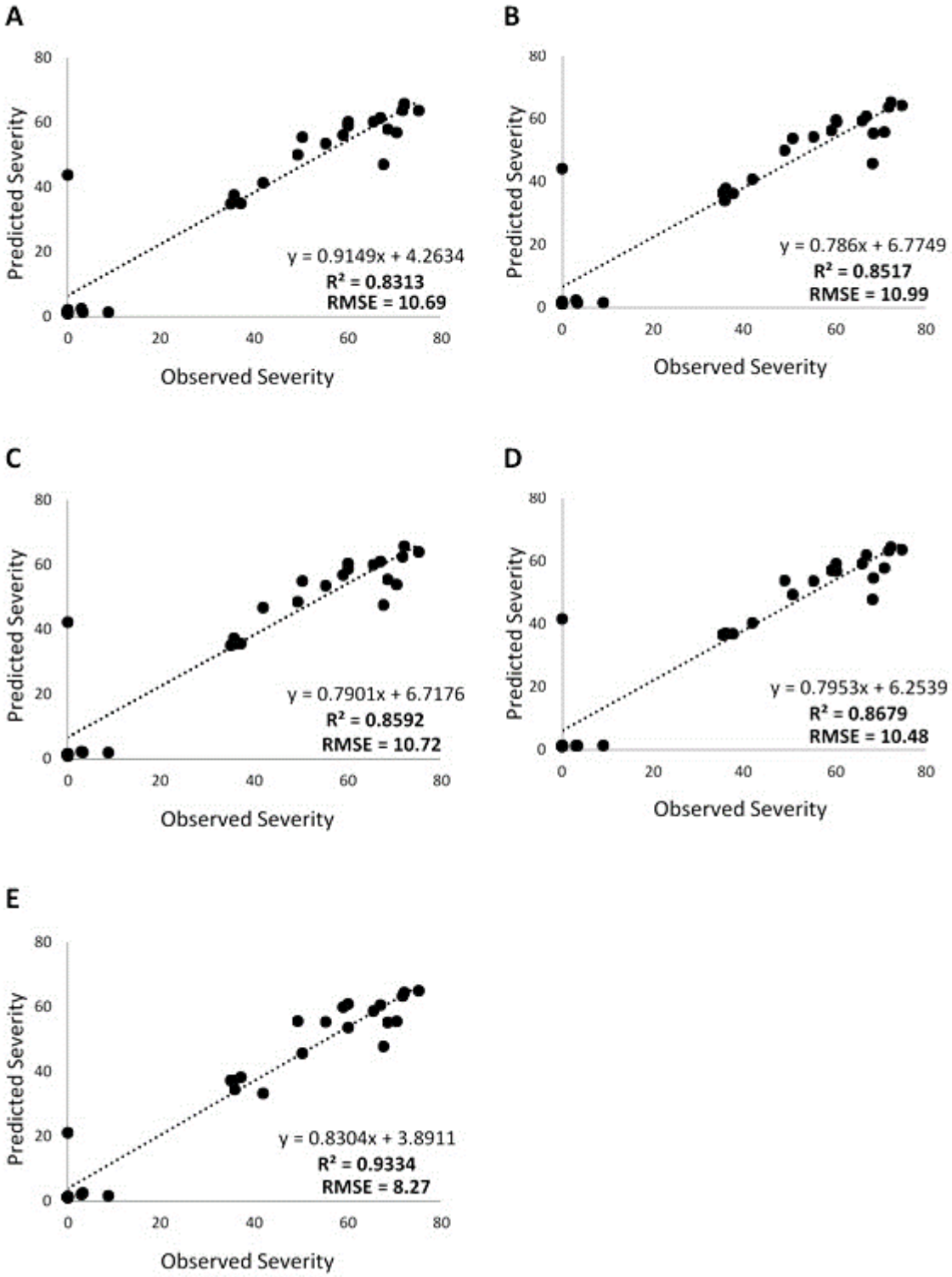

Figure 11: Prediction of late blight severity scores for the test set (data that did not participate in modeling). Models were based on vegetation indexes created using the information of four aerial images taken during the evaluation period and the (A) NDVI > 0.2, (B) NDVI > 0.3, (C) NDVI > 0.4, (D) NDVI > 0.5, and (E) NDVI > 0.6 masks.

Like Method 1, the highest R^2 and the lowest RMSE for the test set was 0.93 and 8.27%, respectively, observed for the NDVI > 0.6 mask. This result points out that the NDVI > 0.6 mask was the most efficient in predicting disease severity scores on

both methods. Moreover, the use of the selected features (NDVI, Soil-adjusted Vegetation Index (SAVI), Green Normalized Difference Vegetation (GNDVI), MCARI1, and Simple Ratio (SR) from four evaluation days) was efficient in predicting disease severity. The lowest R^2 value (0.83) was found for the NDVI<0.2 mask. The highest RMSE value (10.99%) was found for the NDVI<0.3 mask. Overall, Method 2 was better than Method 1 in predicting late blight severity scores as it has the lowest RMSE-CV (7.62%) and highest R^2 (0.86). Also, Method 2 showed better results for the test set as well (RMSE of 8,27% and R^2 of 0.93).

Best results for the NDVI > 0.6 mask, on both methodologies, suggest significant noise on cropped images. Because most noise pixels were not filtered by small NDVI masks, they were confounded with plant pixels, decreasing model accuracy for late-blight prediction.

Figure 12 and Figure 13 show the importance of variables for Methods 1 and 2, respectively, using the best NDVI mask (NDVI >0.6).

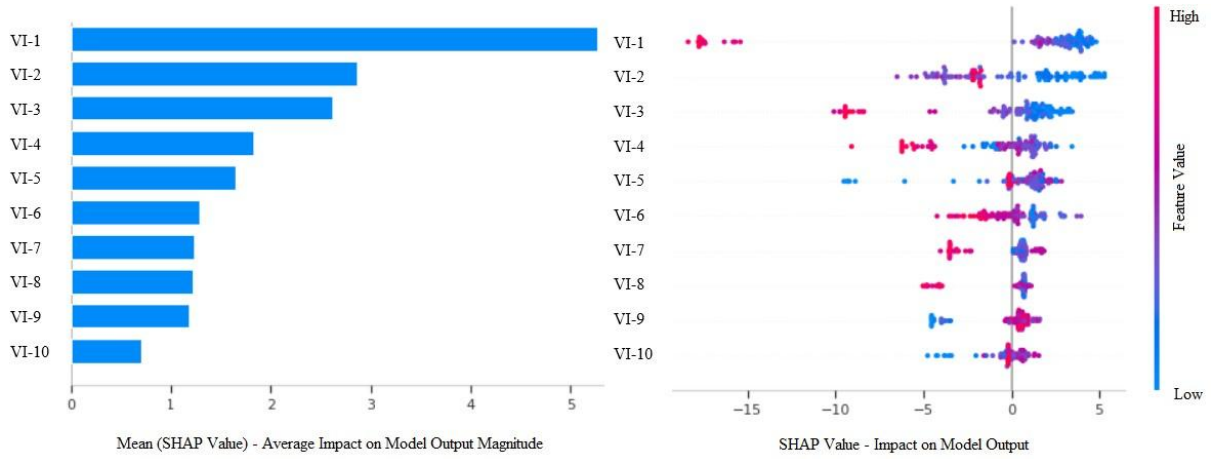

Figure 12: Importance of variables for the model created based on the vegetation indexes (VI) from the last day of evaluation (July 8th) and the NDVI > 0.6 mask (Method 1). VI-1 = SR_d08; VI-2 = MTVI1_d08; VI-3 = MCARI1_d08; VI-4 = NGRDI_d08; VI-5 = NDVIRE_d08; VI-6 = RVI_d08; VI-7 = NDVI_d08; VI-8 = SAVI_d08; VI-9 = NRVI_d08; VI-10 = GNDVI_d08. The last number followed by "d" is the number of the day (d08:08/Jul). Feature value: red color represents observations with high values for the feature, while blue color represents observations with low values.

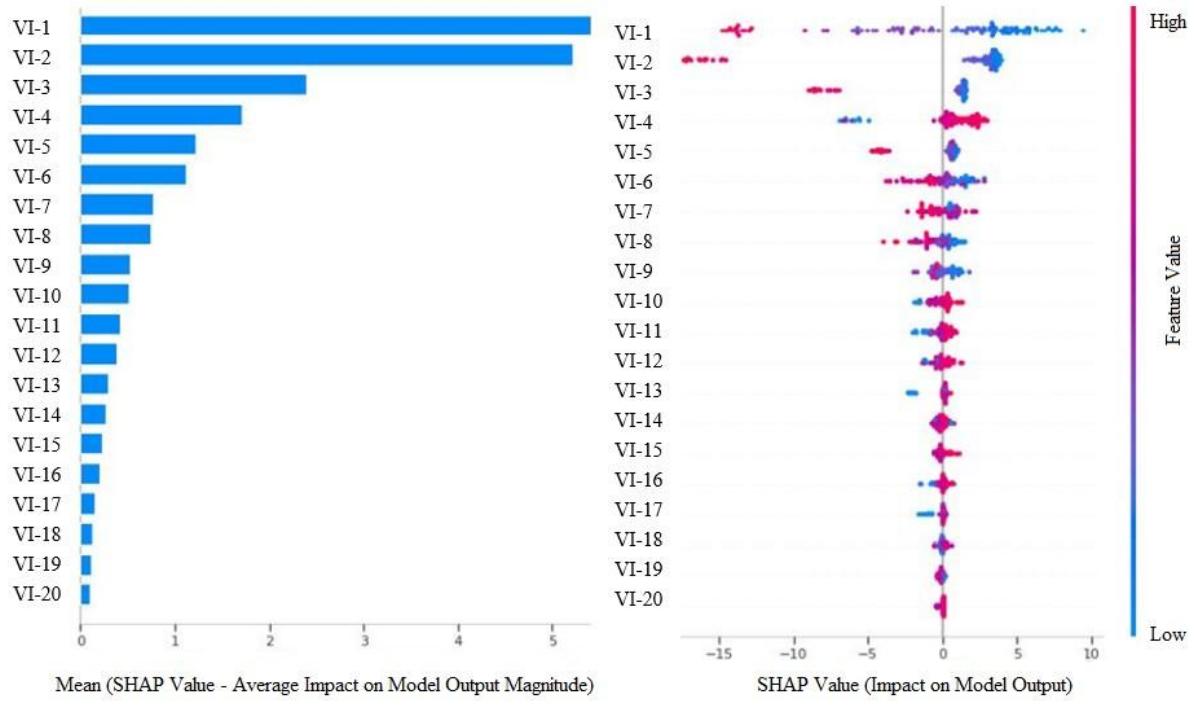

Figure 13: Importance of variables for the model created based on vegetation indexes from all four days of evaluation and the NDVI > 0.6 mask (Method 2). VI-1=MCARI1_d08; VI-2=SR_d08; VI-3=SAVI_d08; VI-4=GNDVI_d28; VI-5=NDVI_d08; VI-6=SR_d04; VI-7=GNDVI_d04; VI-8=MCARI1_d04; VI-9=GNDVI_d08; VI-10=NDVI_d28; VI-11=SR_d01; VI-12=GNDVI_d01; VI-13=MCARI1_d28; VI-14=SAVI_d01; VI-15=NDVI_d01; VI-16=SR_d28; VI-17=SAVI_d04; VI-18=MCARI1_d01; VI-19=NDVI_d04; VI-20=SAVI_d28. The last number followed by "d" is the number of the day (d28: 28/Jun, d01: 01/Jul, d04: 04/Jul, d08: 08/Jul). Feature value: red color represents observations with high values for the feature, while blue color represents observations with low values.

Images on the left (Figure 12 and Figure 13) show SHAP (Shapley Additive Explanations) means for each variable. Based on such values, it is possible to identify the importance of each feature on each model. SR_d08 vegetation index is of great importance to the model in Method 1 (Figure 12). SR_d08 importance value in predicting disease severity in Method 1 was almost twice that of the second most important feature Modified Triangular Vegetation Index 1 (MTVI1_d08). Regarding the model in Method 2 (Figure 13), the MCARI1_d08 was considered the most important variable, followed by SR_d08. Both indexes were calculated based on the aerial images taken on the last day of disease evaluation (July 8th). The three most important features were chosen considering the indexes from the last day of evaluation. After that, the model started to consider the features from the previous evaluation days (June 28th, July 1st and 4th).

Images on the right (Figure 12 and Figure 13) show SHAP distribution values for each variable. The more positive the SHAP value, the higher the disease severity score and vice-versa. Red color represents observations with high values for the feature, while blue color represents observations with low values. High SR_d08 values observed on the right image of Figure 12 contributed to low disease severity scores, which means that plants were healthy. Note that both high and low SR_d08 values are well defined in the positive and negative SHAP sides (Figure 12). A different pattern was observed for NDVIRE_d08 in which there was a small confusion between positive SHAP values and high and low NDVIRE_d08 values. In Figure 13, high values for MCARI1_d08 and SR_d08 indexes resulted in lower disease severity, in other words, the healthier were the plants. Note that, similarly to what was found for Method 1, both high and low SR_d08 values are well defined into the positive and negative SHAP values.

Our data shows that the fittest model for prediction of late blight severity was that using the NDVI > 0.6 filtering mask on Method 2. Genotype ranking based on predicted severity using this model is shown on supplementary file. Slight variations occur between observed severity scores based on the disease severity scale and the predicted severity based on the machine-learning model. Polygons containing the resistant checks were ranked first as they had the lowest severity scores. The model was able to clearly separate resistant checks from susceptible checks (Table 5).

Table 5: Observed and predicted severity for late blight in resistant and susceptible tomato lines.

Lines / checks					
NC1 CEL BR		**NC 25P**		**Santa Clara**	
Obs.	**Pred.**	**Obs.**	**Pred.**	**Obs.**	**Pred.**
0.00	0.21	0.00	0.31	38.38	41.27
0.00	0.47	0.00	0.49	37.83	43.17
0.00	0.53	0.00	0.61	38.15	43.32
0.07	0.62	0.00	0.83	50.05	47.16
0.00	0.95	8.87	1.54	46.07	48.09
0.15	0.96	2.91	2.06	48.07	50.52
0.08	0.99	0.00	2.41	57.48	54.69
0.13	1.03	0.00	5.01	55.99	56.50
1.28	1.13	6.89	5.74	56.30	56.57
3.36	2.47	0.00	5.97	58.24	57.06
11.44	7.38	9.65	6.77	54.39	58.02
8.99	7.90	0.00	21.99	58.99	58.80
Mean					
2.12	2.05	2.36	4.48	50.00	51.26

NC1 CEL BR and NC 25P: resistant checks; Santa Clara: susceptible check; Obs.: observed severity scores expressed in estimates of area under the disease progress curve - AUDPC; Pred.: predicted severity scores based on machine learning models.

Late blight severity scores varied from 0.21 to 21.99 in the resistant checks and from 41.27 to 58.80 in the susceptible checks. Average predicted severity was 2.05, 4.48, and 51.26 for the resistant checks NC1 CELBR and NC 25P, and for the susceptible check Santa Clara, respectively.

5.4. DISCUSSION

Plant breeding efforts consists of phenotyping plants followed by selecting desirable genotypes to continue on the next breeding stages (Singh et al. 2021). In addition, early estimates of pest severity attacks have the potential to reduce control costs decreasing environmental impact (Wspanialy and Moussa 2020). The existing techniques for disease detection are limited by time consumption and evaluation costs of disease severity that are performed visually. Therefore, using multispectral images associated with machine learning techniques could bring quickness, efficiency, and reduction in disease-monitoring costs of large commercial crop fields. Moreover, it could increase genetic gains and suggest changes in disease management strategies (Singh et al. 2021).

Plant reflectance data, obtained from optic sensors on board of an UAV, allows phenotyping of tomato experimental plots. The disease-monitoring approach used

here allowed us to supervise, over time, pathogen development in the inoculated plants. Moreover, through the techniques we applied here, we were also able to predict disease severity scores of tomato plants on the last day of evaluation (worst-case scenario regarding disease development). Such techniques could potentially amplify disease assessment to a larger number of genotypes creating the possibility of high-throughput phenotyping of tomato populations used in breeding programs. The result observed for the R^2 (0.81) and the RMSE (10.34%) on the NDVI > 0.60 mask in Method 1 can be considered satisfactory in a relatively simple model, when we take into account that errors could be linked to the disease evaluation process based solely on visual observations of late blight symptoms, which is rather subjective, besides requiring trained personnel.

Corrêa, Bueno Filho and Carmo (2009) mentioned that the maximum error associated with using a disease diagrammatic scale for visual evaluation of late blight severity scores in tomatoes is about 20%. These authors verified that such errors tend to underestimate the correct late blight severity scores. Considering this error, the prediction of late blight severity using the vegetation indexes together with the NDVI > 0.6 mask in both methodologies showed a good fit.

P. infestans also infects the potato crop. Studies quantifying late blight severity in potato plants using multispectral imaging data and machine learning algorithms are available. Sugiura et al. (2016) observed an R^2 of 0.77 for the regression model between predicted and observed AUDPC in potato plants, whereas Duarte-Carvajalino et al. (2018) observed an average R^2 of 0.75. According to these studies, Method 1 has the advantage of requiring only a single image taken at advanced stages of disease infection. However, we recommend Method 2 to increase accuracy for selection.

Remote sensing approaches involving the use of multispectral cameras detect wave lengths not detected by the human eye that are able to differentiate healthy from unhealthy plant tissues especially at initial stages of infection (Singh et al. 2021). It may explain the low correlation observed between vegetation indexes and observed severity scores in the first disease assessments. The use of image data in tomato phenotyping for disease resistance tend to improve data accuracy by removing the subjectivity of visual scoring (Singh et al. 2021).

In addition, the statistical techniques used in this study are similar to those used in other works. Boechat et al. (2014) found significant inverse correlations between

vegetation indexes and white mold severity. Rumpf et al. (2010) reported that pathogen presence interferes in plant metabolic and physiological processes, causing modifications in plant tissues. Such modifications alter plant spectral patterns and hence are detected by the vegetation indexes. Therefore, we assume that plants had a higher degree of plant tissue modification on the last day of evaluation compared to the previous evaluation days. Hence, higher disease severity scores were detected by the vegetation indexes.

Late blight is a disease that develops quickly causing plant tissues to necrose. Degradation of chlorophyll and other pigments as well as damage to the anatomy of plant organs alter light absorption in the visible and near-infrared light ranges modifying spectral reflectance, allowing us to differentiate sick from healthy plants (Zhang and Qin 2004). Vegetation indexes that best predict late-blight severity are estimated in the green, red, and infra-red band regions (Figure 12 and Figure 13, Table 2). Ray et al. (2011) identified one band in the green, two in the red, and other two in the infrared regions as responsible for accurately differentiating healthy potato plants from *P. infestans*-infected ones. As for Franceschini et al. (2017) three spectral bands performed better. Similarly, in this study, the two vegetation indexes that considered three spectral bands for their estimation showed good prediction results for late blight severity, in both methodologies.

The success of a breeding program depends on the assessment of as many genotypes in early breeding stages as possible (Costa, Bueno Filho and Ramalho 2005). However, such number is frequently limited by the lack of enough resources. To overcome this problem, non-conventional experimental designs, such as replicated check designs or augmented block designs, are used in an attempt to assess a greater number of genetic materials available (Cruz, Carneiro and Regazzi 2014). In this work, while phenotyping tomato lines for *P. infestans* resistance using the disease diagrammatic scale took about four hours, phenotyping using a multispectral camera onboard an UAV took about thirty minutes.

The machine learning model correctly ranked tomato lines based on their disease severity levels since resistant checks occupied the first positions in the rank (supplementary files). The difference observed between mean predicted severity values of the two resistant tomato checks is explained by the number of late blight resistant genes they contain. While NC 25P contains the *Ph-3* gene only, NC 1 CELBR contains both *Ph-2* and *Ph-3* genes, which makes it more resistant. From this

perspective, phenotyping of tomato materials for *P. infestans* resistance through remote sensing and machine learning techniques allows breeders to assess all genotypes available through conventional experimental designs, such as randomized block designs, in early stages of plant breeding.

5.5. CONCLUSION

This study shows that remote sensing and machine learning techniques are reliable phenotyping tools for late blight severity quantification in tomatoes grown at open field conditions. The NDVI > 0.6 was considered the best filtering mask for image processing in the two methodologies tested. Resistant checks had low predicted severity scores by the machine learning model and were classified on top-rank positions confirming model's adequacy on genotype classification. Such approach will allow high throughput phenotyping of tomato genotypes for *P. infestans* resistance at field conditions.

Acknowledgement

Authors would like to thank Jorge Tadeu Fim Rosas for helping to collect the images obtained with unmanned aerial vehicles (UAVs).

Declaration of interest statement

The authors declare no potential conflict of interest.

Funding: This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (Finance Code 001); Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG; and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [FOD], upon reasonable request.

5.6. REFERENCES

Abreu, F.B., D.J.H. Silva, C.D. Cruz, and E.S.G. Mizubuti. 2008. “Inheritance of Resistance to Phytophthora Infestans (Peronosporales, Pythiaceae) in a New Source of Resistance in Tomato (Solanum Sp.(Formerly Lycopersicon Sp.),.” Genetics and Molecular Biology 31 (2). Sociedade Brasileira de Genética: 493–497. doi:10.1590/S1415-47572008000300016.

Alvarenga, M. 2013. Tomate - Produção Em Campo, Casa de Vegetação e Hidropon. 2nd ed. Lavras. https://www.editoraufv.com.br/produto/tomate/1110835.

Baret, F., and G. Guyot. 1991. “Potentials and Limits of Vegetation Indices for LAI and APAR Assessment.” Remote Sensing of Environment 35 (2–3). Elsevier: 161–173. doi:10.1016/0034-4257(91)90009-U.

Bergougnoux, V. 2014. “The History of Tomato: From Domestication to Biopharming.” Biotechnology Advances 32 (1). Elsevier: 170–189. doi:10.1016/J.BIOTECHADV.2013.11.003.

Boechat, L., F.A.C. Pinto, T.J. de Paula, D.M. Queiroz, and H. Teixeira. 2014. “Detecção Do Mofo-Branco No Feijoeiro, Utilizando Características Espectrais.” Revista Ceres 61 (6). Universidade Federal de Viçosa: 907–915. doi:10.1590/0034-737X201461060004.

Brownlee, J. 2019. “How to Choose a Feature Selection Method For Machine Learning.” https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/.

Campbell, C. L., and L. V. Madden. 1990. “Introduction to Plant Disease Epidemiology.” Wiley, 532.

Chawade, A., J. Van Ham, H. Blomquist, O. Bagge, E. Alexandersson, and R. Ortiz. 2019. “High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture.” Agronomy. Multidisciplinary Digital Publishing Institute. doi:10.3390/agronomy9050258.

Corrêa, F.M., J.S.S. Bueno Filho, and M.G.F. Carmo. 2009. “Comparison of Three Diagrammatic Keys for the Quantification of Late Blight in Tomato Leaves.” Plant Pathology 58 (6). Blackwell Publishing Ltd: 1128–1133. doi:10.1111/j.1365-3059.2009.02140.x.

Costa, J., J.S.S. Bueno Filho, and M.A.P. Ramalho. 2005. “Neighborhood and Spatial Analysis in Plant Breeding.” Pesquisa Agropecuaria Brasileira 40 (11): 1073–1079. doi:10.1590/s0100-204x2005001100004.

Cruz, A., Y. Ampatzidis, R. Pierro, A. Materazzi, A. Panattoni, L. De Bellis, and A. Luvisi. 2019. “Detection of Grapevine Yellows Symptoms in Vitis Vinifera L. with Artificial Intelligence.” Computers and Electronics in Agriculture 157 (February). Elsevier: 63–76. doi:10.1016/j.compag.2018.12.028.

Cruz, C.D., P.C.S. Carneiro, and A.J. Regazzi. 2014. Modelos Biométricos Aplicados Ao Melhoramento Genético - Vol 3. 3o. Viçosa: Editora UFV.

Daughtry, C. S.T., C. L. Walthall, M. S. Kim, E. Brown De Colstoun, and J. E. McMurtrey. 2000. “Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance.” Remote Sensing of Environment 74 (2). Elsevier: 229–239. doi:10.1016/S0034-4257(00)00113-9.

Deng, L., Mao, Z., Li, X., Hu, Z., Duan, F., Yan, Y. 2018. "UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras." ISPRS Journal of Photogrammetry and Remote Sensing 146: 124-136. doi.org/10.1016/j.isprsjprs.2018.09.008

DJI Matrice 100, 2021. https://www.dji.com/br/matrice100/info (accessed 27 February 2021)

Duarte, H.S.S, L. Zambolim, and W.C. Jesus Junior. 2007. “Manejo Da Requeima Do Tomateiro Industrial Empregando Sistema de Previsão.” Summa Phytopathologica 33 (4). FapUNIFESP (SciELO): 328–334. doi:10.1590/s0100-54052007000400002.

Duarte-Carvajalino, J. M., Alzate, D. F., Ramirez, A. A., Santa-Sepulveda, J. D., Fajardo-Rojas, A. E., Soto-Suárez, M. 2018. " Evaluating Late Blight Severity in Potato Crops Using Unmanned Aerial Vehicles and Machine Learning Algorithms." Remote Sensing 10(10):1513. doi: 10.3390/rs10101513.

Eshed, Y., and D. Zamir. 1995. “An Introgression Line Population of Lycopersicon Pennellii in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield- Associated QTL.” Genetics 141 (3). Genetics Society of America: 1147–1162. /pmc/articles/PMC1206837/?report=abstract.

Filgueira, F.A.R., P.C. Obeid, H.J. Morais, W.V. Santos, and R.R. Fontes. 1999. “Tomate Tutorado.” In Recomendações Para o Uso de Corretivos e Fertilizantes Em Minas Gerais - 5a Aproximação, edited by A.C. Ribeiro, P.T.G. Guimarães, and V.H.V. Alvarez, 187–188. Viçosa: Editora SBCS.

Franceschini, M.H.D., Bartholomeus, H., van Apeldoorn, D., Suomalainen, J., Kooistra, L. 2017. "Assessing changes in potato canopy caused by late blight in organic production systems through UAV-based pushbroom imaging spectrometer." Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 42, 109–112. doi.org/10.5194/isprs-archives-XLII-2-W6-109-2017.

Furlani, P.R., V. Faquin, M.A.R. Alvarenga, and S. Seno. 2013. “Produção Em Substratos e Em Hidroponia.” In Tomate: Produção Em Campo, Casa de Vegetação e Hidroponia, edited by MAR Alvarenga, 245–273. Lavras: Editora Universitária de Lavras.

Gardner, R.G., and D.R. Panthee. 2010a. “‘Plum Regal’ Fresh-Market Plum Tomato Hybrid and Its Parents, NC 25p and NC 30p.” HortScience 45 (5). American Society for Horticultural Science: 824–825.

Gardner, R.G., and D.R. Panthee. 2010b. “NC 1 CELBR and NC 2 CELBR: Early Blight and Late Blight-Resistant Fresh Market Tomato Breeding Lines.” HortScience 45 (6). American Society for Horticultural Science: 975–976. doi:10.21273/HORTSCI.45.6.975.

Gitelson, A. A., and M. N. Merzlyak. 1997. “Remote Estimation of Chlorophyll Content in Higher Plant Leaves.” International Journal of Remote Sensing 18 (12). Taylor & Francis Group: 2691–2697. doi:10.1080/014311697217558.

Gitelson, Anatoly A., Yoram J. Kaufman, and Mark N. Merzlyak. 1996. “Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS.” Remote Sensing of Environment 58 (3). Elsevier: 289–298. doi:10.1016/S0034-4257(96)00072-7.

Haboudane, D., J.R. Miller, E. Pattey, P.J. Zarco-Tejada, and I.B. Strachan. 2004. “Hyperspectral Vegetation Indices and Novel Algorithms for Predicting Green LAI of Crop Canopies: Modeling and Validation in the Context of Precision Agriculture.” Remote Sensing of Environment 90 (3): 337–352. doi:10.1016/j.rse.2003.12.013.

Jordan, Carl F. 1969. “Derivation of Leaf-Area Index from Quality of Light on the Forest Floor.” Ecology 50 (4). John Wiley & Sons, Ltd: 663–666. doi:10.2307/1936256.

Luvisi, A., Y.G. Ampatzidis, and L. De Bellis. 2016. “Plant Pathology and Information Technology: Opportunity for Management of Disease Outbreak and Applications in Regulation Frameworks.” Sustainability 2016, Vol. 8, Page 831 8 (8). Multidisciplinary Digital Publishing Institute: 831. doi:10.3390/SU8080831.

MicaSense. 2021. “No Title.” https://micasense.com/pt-br/.

Nowicki, M., E.U. Kozik, and M. Foolad. 2013. “Late Blight of Tomato.” Translational Genomics of Crop Breeding 1 (October): 241–265. doi:10.1002/9781118728475.ch13.

Park, T.H., V.G.A.A. Vleeshouwers, R.C.B. Hutten, H.J. Van Eck, E. Van Der Vossen, E. Jacobsen, and R.G.F. Visser. 2005. “High-Resolution Mapping and Analysis of the Resistance Locus Rpi-Abpt against Phytophthora Infestans in Potato.” Molecular Breeding 16 (1). Springer: 33–43. doi:10.1007/s11032-005-1925-z.

Pedregosa, F., G. Varoquaux, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, et al. 2011. “Scikit-Learn: Machine Learning in Python.” Journal of Machine Learning Research 12: 2825–2830. http://scikit-learn.sourceforge.net.

Prasad, A., Mehta, N., Horak, M., Bae, W.D. 2022. "A two-step machine learning approach for crop disease detection using GAN and UAV technology." *Remote Sens. 14(19): 4765. doi:* 10.3390/rs14194765.

QGIS Development Team. 2019. “Welcome to the QGIS Project. Open Source Geospatial Foundation Project.” QGIS Geographic Information System. http://www.qgis.org/.

Ray, S. S., Jain, N., Arora, R. K., Chavan, S., Panigrahy, S. 2011. "Utility of hyperspectral data for potato late blight disease detection." J Indian Soc Remote Sens 39(2): 161 - 169. doi: 10.1007/s12524-011-0094-2.

Rondeaux, Geneviève, Michael Steven, and Frédéric Baret. 1996. “Optimization of Soil-Adjusted Vegetation Indices.” Remote Sensing of Environment 55 (2). Elsevier: 95–107. doi:10.1016/0034-4257(95)00186-7.

Rouse, J. W., R. H. Haas, J. A. Schell, and D. W. Deering. 1974. “Monitoring Vegetation Systems in the Great Plains with ERTS Proceeding.” In Third Earth Reserves Technology Satellite Symposium, Greenbelt: NASA SP-351, 30103017:317. Proceedings, Washington, NASA. https://ui.adsabs.harvard.edu/abs/1974NASSP.351..309R/abstract.

Rumpf, T., A. K. Mahlein, U. Steiner, E. C. Oerke, H. W. Dehne, and L. Plümer. 2010. “Early Detection and Classification of Plant Diseases with Support Vector Machines Based on Hyperspectral Reflectance.” Computers and Electronics in Agriculture 74 (1). Elsevier: 91–99. doi:10.1016/J.COMPAG.2010.06.009.

Shamshiri, R. R., Jones, J. W., Thorp, K. R., Ahmad, D., Man, H. C., Taheri, S. 2018. "Review of optimum temperature, humidity, and vapour pressure deficitfor microclimate evaluation and control in greenhouse cultivation of tomato: a review." Int. Agrophys. 32: 287-302. doi: 10.1515/intag-2017-0005.

Singh, A., S. Jones, B. Ganapathysubramanian, S. Sarkar, D. Mueller, K. Sandhu, and K. Nagasubramanian. 2021. “Challenges and Opportunities in Machine-Augmented Plant Stress Phenotyping.” Trends in Plant Science 26 (1). Elsevier: 53–69. doi:10.1016/J.TPLANTS.2020.07.010.

Sugiura, R., Tsuda, S., Tamiya, S., Itoh, A., Nishiwaki, K., Murakami, N., Shibuya, Y, Hirafuji, M, Nuske, S. 2016. "Field phenotyping system for the assessment of potato late blight resistance using RGB imagery from an unmanned aerial vehicle." Biosystems Engineering 148: 1-10. doi: 10.1016/j.biosystemseng.2016.04.010.

Tucker, Compton J. 1979. “Red and Photographic Infrared Linear Combinations for Monitoring Vegetation.” Remote Sensing of Environment 8 (2). Elsevier: 127–150. doi:10.1016/0034-4257(79)90013-0.

Wspanialy, P., and M. Moussa. 2020. “A Detection and Severity Estimation System for Generic Diseases of Tomato Greenhouse Plants.” Computers and Electronics in Agriculture 178 (November). Elsevier: 105701. doi:10.1016/J.COMPAG.2020.105701.

Xie, C., Y. Shao, X. Li, and Y. He. 2015. “Detection of Early Blight and Late Blight Diseases on Tomato Leaves Using Hyperspectral Imaging.” Scientific Reports 5:1 5 (1). Nature Publishing Group: 1–11. doi:10.1038/srep16564.

Zanotta, S., F.J.S. Salas, J.G. Töfoli, E.S.G. Mizubuti, I.M.L. Terçariol, J.T. Ferrari, R. Domingues, and R. Harakava. 2016. “Requeima: Novos Desafios.” Revista Batata Show 46: 28–32. https://www.researchgate.net/publication/318929301_Requeima_Novos_Desafios_-_Late_Blight_New_Challenges.

Zhang, M., X. Liu, and M. O’Neill. 2002. “Spectral Discrimination of Phytophthora Infestants Infection on Tomatoes Based on Principal Component and Cluster Analyses.” International Journal of Remote Sensing 23 (6). Taylor & Francis Group: 1095–1107. doi:10.1080/01431160110106078.

Zhang, M., Qin, Z. 2004. "Spectral analysis of tomato late blight infections for remote sensing of tomato disease stress in California." International Geoscience and Remote Sensing Symposium VI: 4091-4094. doi: 10.1109/IGARSS.2004.1370031.

5.7. SUPPLEMENTARY FILES

Genotype ranking based on predicted severity estimated through machine learning models. Obs.: observed severity scores expressed in estimates of area under the disease progress curve - AUDPC; Pred.: predicted severity scores based on machine learning models.

Order	Genotype	Disease Severity		Order	Genotype	Disease Severity	
		Obs.	Pred.			Obs.	Pred.
1	NC1CELBR	0.00	0.21	**67**	IL 8-3	60.28	54.68
2	NC25P	0.00	0.31	**68**	Santa Clara	57.48	54.69
3	NC1CELBR	0.00	0.47	**69**	IL 4-1	67.22	54.80
4	NC25P	0.00	0.49	**70**	IL 8-2	54.14	55.27
5	NC1CELBR	0.00	0.53	**71**	IL 4-1-1	59.23	55.51
6	NC25P	0.00	0.61	**72**	Genotype mix	53.76	55.69
7	NC1CELBR	0.07	0.62	**73**	IL 1-1-3	55.63	55.71
8	NC25P	0.00	0.83	**74**	M82	60.50	55.71
9	NC1CELBR	0.00	0.95	**75**	IL 1-2	70.70	55.75
10	NC1CELBR	0.15	0.96	**76**	IL 4-3-2	68.62	55.83
11	NC1CELBR	0.08	0.99	**77**	IL 3-2	52.98	55.87
12	NC1CELBR	0.13	1.03	**78**	IL 2-6	49.25	55.92
13	NC1CELBR	1.28	1.13	**79**	IL 6-4	53.30	56.12
14	NC25P	8.87	1.54	**80**	IL 10-1-1	53.98	56.13
15	NC25P	2.91	2.06	**81**	IL 1-3	55.13	56.17
16	NC25P	0.00	2.41	**82**	IL 4-4	59.44	56.24
17	NC1CELBR	3.36	2.47	**83**	Santa Clara	55.99	56.50
18	NC25P	0.00	5.01	**84**	Santa Clara	56.30	56.57
19	NC25P	6.89	5.74	**85**	IL 3-5	61.60	56.63
20	NC25P	0.00	5.97	**86**	Santa Clara	58.24	57.06
21	NC25P	9.65	6.77	**87**	IL 9-1-2	57.87	57.41
22	NC1CELBR	11.44	7.38	**88**	IL 1-1	53.20	57.43
23	NC1CELBR	8.99	7.90	**89**	IL 10-2	55.61	57.65
24	NC25P	0.00	21.99	**90**	IL 7-2	58.57	57.85
25	LA716	35.72	32.53	**91**	Santa Clara	54.39	58.02
26	LA716	28.48	32.86	**92**	IL 10-2-2	54.48	58.45
27	M82	42.06	33.09	**93**	IL 12-4	59.92	58.73
28	LA716	30.47	33.48	**94**	Santa Clara	58.99	58.80
29	LA716	35.14	35.27	**95**	IL 10-1	63.07	58.88
30	LA716	37.78	35.43	**96**	IL 9-3	59.55	58.92
31	LA716	37.31	36.18	**97**	IL 5-4	60.25	59.05
32	IL 2-1	25.02	36.81	**98**	IL 8-1-1	65.88	59.26
33	LA716	39.60	36.98	**99**	IL 7-4-1	63.67	59.41
34	LA716	35.90	37.29	**100**	IL 5-2	60.37	59.86

35	LA716	36.76	37.65	**101**	IL 5-5	59.16	60.14
36	LA716	29.52	38.39	**102**	IL 4-2	63.77	60.47
37	LA716	36.84	38.83	**103**	IL 9-3-1	60.31	60.57
38	LA716	41.72	39.34	**104**	M82	63.84	60.63
39	Santa Clara	38.38	41.27	**105**	IL 5-3	61.64	60.72
40	IL 2-5	36.93	41.38	**106**	M82	60.21	60.74
41	Santa Clara	37.83	43.17	**107**	M82	68.60	60.86
42	Santa Clara	38.15	43.32	**108**	IL 1-4-18	67.23	61.00
43	IL 1-1-2	39.71	44.41	**109**	IL 10-3	61.20	61.75
44	M82	50.40	44.42	**110**	IL 9-1	62.67	62.27
45	M82	43.62	44.73	**111**	IL 11-2	66.97	62.51
46	M82	46.57	45.43	**112**	IL 8-3-1	61.02	62.64
47	IL 3-3	42.49	45.69	**113**	IL 8-1-3	67.56	62.90
48	M82	68.04	46.97	**114**	IL 9-2-5	63.91	63.08
49	Santa Clara	50.05	47.16	**115**	IL 5-1	67.23	63.40
50	IL 6-3	42.16	47.91	**116**	IL 9-2	64.14	63.57
51	Santa Clara	46.07	48.09	**117**	IL 9-1-3	62.16	63.65
52	IL 3-1	52.18	48.42	**118**	IL 11-4	71.95	63.65
53	IL 2-6-5	48.86	49.69	**119**	IL 12-3	72.32	64.06
54	IL 7-1	43.63	49.83	**120**	IL 7-5	62.86	64.08
55	IL 1-4	51.99	50.45	**121**	IL 11-3	68.63	64.30
56	Santa Clara	48.07	50.52	**122**	IL 6-2	65.50	64.30
57	IL 2-4	49.69	50.66	**123**	IL 7-5-5	70.09	65.08
58	IL 2-1-1	45.17	50.75	**124**	IL 11-4-1	64.78	65.15
59	IL 8-1	54.07	52.15	**125**	M82	67.34	65.33
60	IL 3-4	56.61	52.44	**126**	IL 8-2-1	75.10	65.44
61	IL 12-4-1	49.79	52.46	**127**	IL 9-3-2	67.77	66.30
62	IL 4-3	64.19	52.73	**128**	IL 12-1	68.65	66.45
63	M82	53.80	53.06	**129**	IL 12-2	71.17	66.72
64	IL 6-1	51.27	53.36	**130**	IL 7-4	69.54	67.16
65	M82	63.68	53.37	**131**	IL 12-3-1	67.89	67.40
66	IL 2-3	50.93	53.80	**132**	IL 12-1-1	68.73	68.30