CÉSAR PEDRO

USO DE APRENDIZADO DE MÁQUINA PARA OTIMIZAÇÃO DA SELEÇÃO RECORRENTE RECÍPROCA EM MILHO

LAVRAS – MG

2025

CÉSAR PEDRO

USO DE APRENDIZADO DE MÁQUINA PARA OTIMIZAÇÃO DA SELEÇÃO RECORRENTE RECÍPROCA EM MILHO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do programa de Pós-Graduação em Genética e Melhoramanto de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Orientador

Prof. DSc. João Cândido de Souza

LAVRAS – MG

2025

Ficha de identificação da obra elaborada pelo(a) autor(a) através do Sistema de Geração Automática de Ficha Catalográfica da Biblioteca Universitária da UFLA.

Pedro, César .
Uso de aprendizado de máquina para otimização da seleção recorrente recíproca em milho / César Pedro. 2025.
76 p. : il.

Orientador: João Cândido de Souza

Tese (Doutorado) - Universidade Federal de Lavras, 2025.
Bibliografia.

1. Zea mays L. 2. Índice de seleção. 3. Análise de fatores. 4. Modelos multiníveis. 5. Fenotipagem. I. Cândido de Souza, João . II. Universidade Federal de Lavras. III. Título.

CÉSAR PEDRO

USO DE APRENDIZADO DE MÁQUINA PARA OTIMIZAÇÃO DA SELEÇÃO RECORRENTE RECÍPROCA EM MILHO

APPLICATION OF MACHINE LEARNING FOR OPTIMIZING RECIPROCAL RECURRENT SELECTION IN MAIZE

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramanto de Plantas, para a obtenção do título de Doutor.

APROVADO em 30 de Janeiro de 2025.

DSc. João Cândido de Souza UFLA

DSc. Vinícius Quintão Carneiro UFLA

DSc. Marcela Pedroso Mendes Externo

DSc. Livia Maria Chamma Davide Externo

DSc. Deoclécio Domingos Garbuglio Externo

gov.br
Documento assinado digitalmente
JOAO CANDIDO DE SOUZA
Data: 16/04/2025 14:42:11-0300
Verifique em https://validar.iti.gov.br

Orientador

Prof. DSc. João Cândido de Souza

LAVRAS – MG

2025

Dedico a minha esposa Delénia Margarida Rodrigues Rico, aos meus filhos Norman César Pedro, as minhas filhas Atia César Pedro e Maria César Pedro. Aos meus pais Pedro Cuhia e Atia Bernardo. A todos meus irmãos.

AGRADECIMENTOS

Agradeço à Universidade Federal de Lavras – UFLA, ao Departamento de Biologia, ao Programa de Pós- Graduação em Genética e Melhoramento de Plantas da UFLA, pela estrutura oferecida e pela oportunidade de realização do Doutorado.

Agradeço a Capes, CNPq e FAPEMIG, pela concessão da bolsa de estudos e recursos, que tornou possível a realização da pesquisa de Doutorado.

RESUMO

A Seleção Recorrente Recíproca (SRR) tem sido amplamente empregada no melhoramento genético do milho, com o objetivo de maximizar o potencial genético e a heterose entre populações. Contudo, sua eficácia está intrinsecamente ligada à análise precisa de caracteres e progênies de milho. Nesse contexto, metodologias baseadas em aprendizado de máquina e modelos estatísticos avançados podem ser conjuntamente utilizadas para otimizar a seleção e predizer o desempenho em programas de SRR. Este trabalho teve como objetivo principal otimizar o processo de seleção de progênies de milho no programa de SRR da Universidade Federal de Lavras, por meio do uso de técnicas de aprendizado de máquina. Os experimentos foram conduzidos na Universidade Federal de Lavras entre os anos de 2021 a 2024, em um delineamento Alpha Lattice. O estudo foi estruturado em quatro etapas principais: (i) análise do potencial genético das progênies por meio de estratégias baseadas na integração da metodologia Best Subset Regression para otimização da seleção simultânea de características por meio índices de seleção. (ii) recombinação de 15% das progênies selecionadas, seguida da geração de quatro populações de intracruzamento e intercruzamento pelo metodo Bulk Pollen Pollination. Essas populações foram avaliadas e submetidas as análises de parâmetros genéticos e fatoriais baseadas em correlação fenotípica e genética, visando seleção simulatanea de progenies com expressão fenotípica e genética de suas características e otimização por agrupamento de K-means. (iii) Predição da produtividade das progênies por meio de modelos multiníveis no contexto de melhoramento aninhado (progenies dentro de populações e estas dentro cruzamentos), e (iv) integração da fenotipagem de alto rendimento a modelos de aprendizado de máquina baseados em árvores interpretáveis e índices de seleção simultânea. Em conclusão, a integração das metodologias Best Subset Regression e índice de seleção simultânea é uma abordagem eficaz para orientar a identificação de linhagens promissoras e impulsionar o progresso genético das populações de milho em SRR. As estratégias de seleção fatorial em populações mostraram eficácia ao identificar progênies com desempenho simultâneo em relação à expressão fenotípica e genética, contribuindo para o desenvolvimento de cultivares de maior estabilidade agronômica e genética. O modelo multinível destacou-se como ferramenta com maior poder preditivo e suporte à seleção acurada de progênies superiores. A integração dos modelos de aprendizado de máquina com índice de seleção simultânea maximizou a resposta de seleção por meio da escolha de progênies de milho com ótimos ganhos genéticos. O estudo valida o aprendizado de máquina e sua integração com modelos estatísticos como uma estratégia inovadora e eficaz para a otimização da SRR em milho, acelerando o progresso do melhoramento genético e contribuindo para o desenvolvimento de híbridos de elevado desempenho.

Palavras-chave: *Zea mays* L.; Índice de seleção; Análise de fatores; Modelos multiníveis; Fenotipagem.

ABSTRACT

Reciprocal Recurrent Selection (RRS) has been widely used in maize breeding to maximize genetic potential and heterosis between populations however, its effectiveness is intrinsically linked to the precise analysis of maize traits and progenies. In this context, machine learning-based methodologies and advanced statistical models can be jointly employed to optimize selection and predict performance in RRS programs. The main objective of this study was to optimize the selection process of maize progenies in the Federal University of Lavras RRS program using machine learning techniques. The experiments were conducted at the Federal University of Lavras between 2021 and 2024 using an Alpha Lattice design. The study was structured into four main stages: i) Analysis of the genetic potential of progenies through strategies based on the integration of the Best Subset Regression methodology to optimize the simultaneous selection of traits using selection indices; ii) recombination of the top 15% of selected progenies followed by the generation of four intrapopulation and interpopulation crosses using the Bulk Pollen Pollination method. These populations were evaluated and subjected to genetic and factorial parameter analyses based on phenotypic and genetic correlations aiming for the simultaneous selection of progenies with optimal phenotypic and genetic expression of their traits, optimized through K-means clustering. iii) Prediction of progeny yield using multilevel models in the context of nested breeding progenies within populations and populations within crosses. iv) Integrating high-throughput phenotyping with interpretable tree-based machine learning models and simultaneous selection indices. In conclusion, the integration of Best Subset Regression and simultaneous selection indices proved to be an effective approach for identifying promising lineages and enhancing the genetic progress of maize populations in RRS. Factorial selection strategies in populations demonstrated efficacy in identifying progenies with simultaneous phenotypic and genetic performance contributing to the development of cultivars with greater agronomic and genetic stability. The multilevel model stood out as a tool with high predictive power, supporting the accurate selection of superior progenies. The integration of machine learning models with simultaneous selection indices maximized selection response by choosing maize progenies with optimal genetic gains. This study validates machine learning and its integration with statistical models as an innovative and effective strategy for optimizing RRS in maize, accelerating genetic breeding progress and contributing to the development of high-performance hybrids.

Keywords: *Zea mays* L.; Selection index; Factor analysis; Multilevel models; Phenotyping.

INDICADORES DE IMPACTO

O estudo demonstra que a integração de Machine Learning (ML) e modelos estatísticos avançados na Seleção Recíproca Recorrente (SRR) de milho otimiza a identificação de progênies superiores, acelerando o desenvolvimento de híbridos de alto desempenho. No aspecto social, a metodologia beneficia agricultores, especialmente pequenos e médios produtores, ao aumentar a disponibilidade de sementes mais produtivas e adaptáveis, reduzindo a insegurança alimentar e fortalecendo a agricultura familiar em regiões como o Cerrado e o Sudeste de Minas Gerais. No plano econômico, a técnica eleva a eficiência do melhoramento genético, reduzindo custos e tempo de seleção, com potencial de aumentar a produtividade em mais de 20%, impulsionando a competitividade do milho mineiro e do Brasil. Ambientalmente, a seleção de cultivares mais produtivas diminui a necessidade de expansão de áreas de cultivo e, consequentemente, aumenta a eficiência do uso de recursos naturais, alinhando-se aos ODS 2 (Fome Zero e Agricultura Sustentável) e ODS 13 (Ação Contra a Mudança Global do Clima). Do ponto de vista tecnológico, a aplicação de ML e modelos multiníveis moderniza programas de melhoramento, posicionando a pesquisa regional e nacional na vanguarda da agricultura moderna. Os impactos enquadram-se na área temática 7 (Tecnologia e Produção) da Política Nacional de Extensão, com contribuições diretas para inovação agrícola sustentável e segurança alimentar, em sintonia com a Agenda 2030 da ONU. A abordagem mostra-se uma estratégia replicável em outras culturas, ampliando seu potencial transformador para a agricultura tropical.

IMPACT INDICATORS

The study demonstrates that integrating Machine Learning (ML) and advanced statistical models into Reciprocal Recurrent Selection (RRS) for maize optimizes the identification of superior progenies, accelerating the development of high-performance hybrids. From a social perspective, this methodology benefits farmers, particularly small and medium-sized producers, by increasing the availability of more productive and adaptable seeds. This reduces food insecurity and strengthens family farming in regions such as the Cerrado and Southeastern Minas Gerais. Economically, the technique enhances genetic breeding efficiency, reducing selection costs and time, with the potential to increase productivity by over 20%. This boosts the competitiveness of Minas Gerais maize and Brazil's overall production. Environmentally, selecting higher-yielding cultivars reduces the need for cropland expansion, consequently improving natural resource use efficiency. This aligns with SDG 2 (Zero Hunger and Sustainable Agriculture) and SDG 13 (Climate Action). From a technological standpoint, applying ML and multilevel models modernizes breeding programs, positioning regional and national research at the forefront of modern agriculture. The impacts fall under Thematic Area 7 (Technology and Production) of Brazil's National Extension Policy, contributing directly to sustainable agricultural innovation and food security, in line with the UN 2030 Agenda. The approach proves to be a replicable strategy for other crops, expanding its transformative potential for tropical agriculture.

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

O milho (*Zea mays* L.) é um dos pilares da segurança alimentar global, demandando constantes avanços em produtividade e sustentabilidade. Nesse cenário, a Seleção Recorrente Recíproca (SRR) destaca-se como uma das estratégias mais eficazes para o melhoramento genético contínuo, promovendo ganhos simultâneos na capacidade combinatória de populações distintas e maximizando o potencial agronômico das progênies (Hallauer *et al*., 2010; Combs; Bernardo, 2013). No entanto, o sucesso desse método depende não apenas da variabilidade genética disponível, mas também da identificação precisa de combinações ótimas de características que impulsionem a resposta à seleção, especialmente em cenários onde múltiplas variáveis interagem de forma complexa.

A SRR opera sob o princípio da seleção e recombinação cíclica de duas populações geneticamente distintas, explorando a heterose e preservando a diversidade genética (Combs & Bernardo, 2013). Contudo, um dos principais desafios reside na seleção simultânea de múltiplas características, uma vez que a inclusão indiscriminada de todos os traços avaliados pode resultar em ganhos subótimos devido a redundâncias que podem existir entre as características.

Métodos tradicionais, como os índices de Smith (1936) e Hazel (1943), ou abordagens mais recentes como FAI-BLUP (Rocha *et al*., 2018) e MGIDI (Olivoto; Nardino, 2021), oferecem soluções ao agregar diferentes características em uma única métrica. No entanto, essas técnicas não exploram de forma eficiente quais subconjuntos de características são mais impactantes para maximizar o ganho genético.

Diante dessa lacuna, este estudo propõe uma abordagem inovadora que integra regressão por melhores subconjuntos (Best Subset Regression) com índices de seleção simultânea, visando identificar as combinações ótimas de características que maximizam os ganhos genéticos por seleção. Além disso, investiga-se a aplicação de análises fatoriais baseadas em correlação genética e fenotípica para aprimorar a seleção de progênies de alto desempenho com expressão fenotípica e genética simultaneamente. Na sequência, aplicou-se a abordagem de modelos multiníveis (MLM), capazes de capturar a estrutura hierárquica do contexto de melhoramento (cruzamentos, populações e progênies), visando explicar a

variabilidade e predizer com precisao a produtividade de progênies de milho. Outro avanço incorporado neste trabalho é a integração entre fenotipagem de alta precisão, aprendizado de máquina (AM) e índices de seleção multitrait. A fenotipagem digital de espigas, associada a algoritmos preditivos baseados em árvores interpretáveis, permite a triagem eficiente de características morfológicas relevantes, reduzindo redundâncias e aumentando a acurácia da seleção. Essa sinergia metodológica não apenas otimiza a identificação de progênies superiores, mas também traduz dados complexos em decisões práticas de melhoramento, garantindo maior eficiência no desenvolvimento de cultivares produtivas.

A justificativa para este estudo reside na necessidade de abordagens mais precisas e eficientes para a SRR, capazes de superar os desafios da seleção multitrait e da predição de produtividade em cenários de grande complexidade de dados. Os resultados esperados incluem: a identificação dos subconjuntos de características que maximizam o ganho genético em milho; a validação da superioridade de modelos multiníveis e de AM na predição de produtividade; a seleção de progênies de alto desempenho com base em correlações genéticas e fenotípicas; um framework integrado que combine fenotipagem digital, AM e índices de seleção para otimizar programas de SRR.

Em suma, esta pesquisa não apenas avança o conhecimento teórico em genética quantitativa, mas também oferece ferramentas práticas para o melhoramento de milho. Ao unir metodologias de aprendizado de máquina e estatísticas, o estudo estabelece um novo paradigma de otimização da selecao recorrente reciproca de progênies de milho com extensão para outras culturas agrícolas.

REFERÊNCIAS

COMBS, E.; BERNARDO, R. **Genome-wide selection to introgress semiexotic maize germplasm into U.S. corn belt inbreds**. Crop Science, v. 53, n. 4, p. 1427-1436, 2013.

HALLAUER, A. R.; CARENA, M. J.; MIRANDA FILHO, J. B. **Quantitative Genetics in Maize Breeding**. 3. ed. New York: Springer, 2010.

HAZEL, L. N. **The genetic basis for constructing selection indexes**. Genetics, v. 28, n. 6, p. 476-490, 1943.

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

ROCHA, J. R.; MACHADO, J. C.; CARNEIRO, P. C. S. FAI-BLUP: **A new approach for efficient genomic selection**. G3: Genes, Genomes, Genetics, v. 8, n. 7, p. 2465-2476, 2018.

SMITH, H. F. **A discriminant function for plant selection**. Annals of Eugenics, v. 7, n. 3, p. 240-250, 1936.

SEGUNDA PARTE-ARTIGOS

ARTIGO 1- Multi-trait reciprocal recurrent selection strategies based on best subset regression in maize

Periódico: Euphytica, versão preliminar.

César Pedro (https://orcid.org/0000-0002-2963-8652) and João Cândido de Souza a* (https://orcid.org/0000-0001-9580-4631)

Universidade Federal de Lavras, Departamento de Biologia, Aquenta Sol, Lavras - MG, CEP 37200-900, Brasil.

*e-mail: cansouza@ufla.br (corresponding author)

Abstract: This study aimed to identify the best model for combining traits via Best Subset Regression (BSR) and the simultaneous selection index (SSI) to maximize genetic gains in reciprocal recurrent selection (SRR) of maize multi-trait progenies. The experiment evaluated 56 and 46 interpopulation hybrids (PAB and PBA, respectively) derived from reciprocal crosses of the PA and PB populations and two checks, in an alpha-lattice design, at the Center for Scientific and Technological Development of Agriculture of the Federal University of Lavras/Brazil. Genetic parameters were estimated by Restricted Maximum Likelihood (REML) and means by Best Linear Unbiased Prediction (BLUP). The best trait combination models (M) were defined via BSR and used in SSI (Factor analysis and ideotype-design: FAI-BLUP, Multi-trait Genotype-Ideotype Distance: MGIDI, and Smith-Hazel), with 20% selection intensity. The simultaneous selection efficiency (SSE%) was evaluated in comparison to direct selection (DIS) and the full model (FM), in addition to the agreement between the hybrids selected by the different selection strategies. The results showed genetic variability in both populations and combinations of BSR and SSI that maximized SSE%. The M2 + MGIDI and M3 + MGIDI models of the GY+PROL+SM and GY+PROL+SM+ASI trait combinations provided SSE% of 99.33 and 98.92 for PAB and PBA near DIS and 1.01 to 1.00 of SSE% over FM, respectively. Integrating BSR and SSI methodologies is an effective approach to guide the identification of promising progenies and boost the genetic progress of maize populations in SRR.

Keywords *Zea mays* L . BLUP . Selection indices.

Resumo: objectivou-se identificar o melhor modelo de combinação das caracteristicas via Best Subset regression (BSR) e o índice de seleção simultânea (ISS) para maximizar os ganhos genéticos em seleção recorrente recíproca (SRR) de linhagens multtrait de milho. O experimento, avaliou 56 e 46 híbridos interpopulacionais (PAB e PBA, respectivamante) derivadas de cruzamentos recíprocos das populações PA e PB, e duas testemunhas, em delineamento alfa-látice, no Centro de Desenvolvimento Científico e Tecnológico da Agricultura da Universidade Federal de Lavras/Brasil. Os parâmetros genéticos foram estimados por Restricted Maximum Likelihood (REML) e as medias por Best Linear Unbiased Prediction (BLUP). Os melhores modelos de combinacoes das caracteristicas (M) foram definidos via BSR e utilizados nos ISS (Factor analysis and ideotype-design: FAI-BLUP, Multitrait Genotype-Ideotype Distance: MGIDI, e Smith-Hazel), com 20% de intensidade de seleção. Avaliaram-se a eficiência de seleção simultânea (ESS%) em relação à seleção direta (DS) e ao modelo completo (FM), além da concordância entre os hibridos selecionados pelas diferentes estrategias de selecao. Os resultados mostraram variabilidade genética em ambas as populações e combinações de BSR e ISS que maximizaram a ESS%. Os modelos M2 + MGIDI e M3 + MGIDI das combinacoes das caracteristicas GY+PROL+SM e GY+PROL+SM+ASI, proporcionaram ESS% de 99.33 e 98.92 para PAB e PBA proximo a DS, e 1.01 a 1 de ESS% sobre o FM. A integração das metodologias BSR e ISS é uma abordagem eficaz para orientar na identificacao de linhagens promissoras e impulsionar o progresso genético das populacoes de milho em SRR.

Palavras-chave: *Zea mays* L . BLUP . Índices de seleção.

Introduction

Reciprocal recurrent selection (RRS) is an efficient strategy for continuously improving maize populations. It is based on the selection and cyclical recombination of two genetically distinct populations, aiming to simultaneously enhance their combining capacity and maximize genetic gain for agronomic traits of interest. In addition to exploiting heterosis by strengthening complementarity between populations, RRS boosts yield, preserves genetic variability and generates improved sources for obtaining elite lines, consolidating itself as an essential tool in developing high-performance cultivars.

In RRS, the simultaneous selection of multiple traits is challenging due to the complexity of the interactions between traits and the difficulty in identifying which combinations optimize the response to selection. In scenarios where many traits are evaluated, the simultaneous selection of all available traits may not capture the maximum possible gain, which may result in suboptimal selection of progenies for genetic progress.

Traditionally, selection indices such as Smith (1936) and Hazel (1943), FAI-BLUP Rocha et al. (2018) and MGIDI (Olivoto and Nardino 2021) have been used to combine multiple traits into a single metric, allowing the simultaneous selection of different traits. However, these methods do not efficiently address the selection of the most impactful subsets among a large number of available traits. This study proposes the combination of the best subsets regression methodology (Brooks and Ruengvirayudh 2016) with FAI-BLUP, MGIDI and Smith and Hazel selection indices, aiming to identify the best combinations of traits that maximize the genetic response more effectively. In addition, it compares these combinations with direct selection strategies and those based on all traits with significant genetic variability. Best Subset Regression is an approach that evaluates all possible combinations of traits in a multiple linear regression model, intending to identify the subset that best explains the response trait. This technique allows the selection of the most appropriate model, eliminating irrelevant traits and improving the interpretability of the results (Brooks and Ruengvirayudh 2016). To date, few studies have used this methodology to predict the maize yield and other crops, with emphasis on Zhang et al. (2023), who applied the technique to assess soil health in wheat-maize rotation systems. Paul and Munkvold (2005) combined the best subsets of traits with artificial neural networks to predict the severity of cercospora leaf spot in maize. However, no studies have been carried out in the area of plant breeding that seek to identify the best trait combination model through Best Subset Regression. combined with simultaneous selection indices, aiming to maximize genetic gains in yield and desired response for secondary traits, in RRS programs of maize. The main contribution of this approach to plant breeding is to present an innovative alternative for simultaneous trait selection in maize, integrating regression methodologies and simultaneous selection indices, which assist in the process of selection and combination of relevant traits. This allows the selection of high-performance multitrait progenies for the RRS program and the development of high-yielding varieties.

Materials and Methods

The experiment was conducted between November 2021 and March 2022 in the experimental area of Muquém Farm (21°11'56.9"S 44°58'48.1"W, elevation 918.84 m), which belongs to the Scientific and Technological Development Center for Agriculture at the Federal University of Lavras, located in the southern region of Minas Gerais, Brazil. The soil in the experimental area is classified as Latossolos (Oxisols), according to Santos et al. (2018). The regional climate is humid temperate (Cwa), characterized by dry winters and rainy summers. During the experimental period, the average temperature and precipitation were 23.74°C and 232.56 mm, respectively, as recorded by the Lavras meteorological station (code: 83687, geographic coordinates: 21°13'34.0"S 44°58'47.0"W).

A total of 102 full-sib maize progenies (interpopulation hybrids) were generated by crossing two populations in the eighth cycle of the Reciprocal Recurrent Selection (RRS) program at the Federal University of Lavras. This program, initiated in 2003, used two commercial single-cross hybrids as base populations: DKB 333B (PA) and DOW 657 (PB). The populations were formed by random crosses of 3.000 F1 plants from each single-cross hybrid, resulting in populations in Hardy-Weinberg equilibrium.

The 102 progenies, along with two commercial double-cross hybrids (checks), were evaluated in an alpha-lattice design with three replications and 24 blocks. Sowing was carried out on November 16, 2021, at a density of four seeds per linear meter in plots four meters long, spaced 0.6 meters apart. Basal fertilization consisted of 250 kg ha^{-1} of NPK 8-28-16, while topdressing fertilization, applied 25 days after planting, used 200 kg ha^{-1} of urea (45% N). Other crop management practices followed regional recommendations, as described by Borém et al. (2017).

The evaluated traits included grain yield (GY, kg ha^{-1}), seed mass (SM, g), prolificacy (PROL), days to anthesis (DA), days to silking (DS), anthesis-silking interval (ASI), ear height (EH), and plant height (PH).

The analysis was conducted for an experiment with an alpha-lattice design according to the following model (Resende, 2016):

$$y = Xr + Zg + Wb + e,$$

where: y is the vector of observations, r is the vector of repetition fixed effects, g, b and e are the vectors of random effects of full-sib maize progenies, blocks and errors, respectively. X, Z and W represent the incidence matrices for r, g and b, respectively. The significance of the effects of random progenies from the deviance analysis was verified by the likelihood ratio test (LRT) at 0.001, 0.01 and 0.05 probability.

The variance components were estimated according to Resende (2016) using the Restricted Maximum Likelihood (REML) method, and the means of the effects predicted by Best Linear Unbiased Predictor (BLUP).

The mean heritability and selective accuracy parameters were estimated according to Olivoto and Lúcio (2020). Descriptive statistics (mean, minimum and maximum) were also calculated.

Phenotypic and genetic correlations

The phenotypic and genetic correlations were estimated between the traits of the PAB and PBA populations according to Olivoto and Lúcio (2020). The significance of the correlations was assessed using the t-test at 0.05 probability level.

The best subset regression technique (Brooks and Ruengvirayudh, 2016) was applied to the traits that exhibited genetic variability, aiming to identify the subset of traits that best explains the yield of the populations.

We used $2^p - 1$ possible subsets of combinations of three PAB traits and four PBA traits, resulting in 7 and 15 models, respectively. For each population, the top three and four best trait combination models were selected based on the following criteria: Mallows' C_p: To penalize more complex models, selecting the one with the lowest C_p value. Bayesian Information Criterion (BIC): To penalize complex models more strongly than AIC, choosing the model with the lowest BIC. Adjusted R^2: To adjust the R^2 value for the number of traits in the model, selecting the model with the highest adjusted R^2.

The best trait combinations, along with grain yield, were associated with three simultaneous selection index strategies: FAI-BLUP (Rocha et al. 2018), MGIDI (Olivoto et al. 2019), and Smith (1936) & Hazel (1943). For each strategy, the goal was to increase GY, PROL, and SM while reducing ASI and DS. For the FAI-BLUP and MGIDI indices, the desired ideotype was Max (for increase) and Min (for reduction).

The simultaneous selection efficiency (SSE %) obtained from the relationship between the simultaneous selection gain of the traits (SSG) of the best combination model and the direct selection (DIS) and the full model (FM) for each population was calculated according to the following equation:

$$\text{SSE (\%)} = \left(\frac{\text{SSG}}{\text{DIS}}\right) \times 100; \ \text{SSE (\%)} = \left(\frac{\text{SGG}}{\text{FM}}\right) \times 100$$

The best combination of yield traits and selection strategy was defined by the criterion of the highest ESS% and balance among the associated traits in each population, considering a selection intensity of 20% of the progenies. These results were visually represented using radar charts. Additionally, an analysis of the overlap between the progenies selected by the best strategies and combinations, compared to direct selection, was performed using Venn diagrams. All analyses were conducted in the R software (R Core Team 2024).

Results and discussion

Table 1 presents the genetic and phenotypic parameters of agronomic traits in two maize populations: PAB and PBA. These results aim to evaluate the potential of both populations for the interpopulational breeding program.

The results show that PAB presents significant genetic variability for grain yield (GY), prolificacy (PROL), 100-seed weight (SM), and days to silking (DS). In PBA, significant genetic variability was detected for PROL, SM, DS, and the anthesis-silking interval (ASI). The heritability for GY was moderate in PAB, indicating potential for significant genetic gain with selection. In contrast, in PBA, the lower heritability suggests a limited

response to direct selection for this trait. However, SM and DS stood out in both populations with high heritability and selection accuracy.

Traits such as plant height (PH) presented moderate heritability in both populations. Regarding ear height (EH), heritability was high in PAB but low in PBA. However, genetic variability for PH and EH was not significant, limiting the potential of these traits for selection. For PROL, heritability and accuracy were moderate in both populations, indicating good potential for improvement of this trait and other correlated traits.

The results indicate important differences in the genetic potential of the populations, which can be explored in interpopulation breeding programs. Previous studies, such as those by Reis et al. (2009) and Almada et al. (2024), highlight the relevance of traits with moderate to high genetic variability and high heritability to maximize genetic gains through selection.

GY showed limitations in PBA due to low heritability. However, indirect selection based on correlated traits, such as PROL, SM, and DS, can be an effective strategy. On the other hand, PAB showed better prospects for direct selection of GY, reinforcing its usefulness in reciprocal recurrent selection programs. In this context, PROL, SM, and DS combine significant genetic variability with high heritability and accuracy, essential properties for the indirect selection of high-performance progenies. These advantages are in line with the concept of multi-trait selection, as explained by Cruz et al. (2012) and Olivoto and Nardino (2021).

In addition, these traits offer additional benefits, such as the possibility of selecting progenies with an adequate number of ears per plant, ideal seed mass, and adequate flowering cycle, maximizing genetic gains in yield. The flowering cycle, for example, can be used as a strategic criterion for planning synchronization in interpopulation crosses.

The non-significant genetic variability associated with high heritability in PH and EH can be attributed to experimental precision and not to the genetic variability of progenies, according to Cruz et al. (2012). The presence of significant genetic variability for ASI in PBA demonstrates the potential of this population to recommend progenies in crosses targeting drought-tolerant hybrids. According to Li et al. (2023), ASI is widely used as an indicator in maize progeny selection experiments for drought tolerance, with lower values being desirable for improving yield under stress conditions. However, the low heritability observed in ASI indicates limitations for direct selection, since this trait is strongly influenced by environmental factors, as also pointed out by Li et al. (2023). For ASI, an effective strategy would be indirect selection based on correlated traits with higher heritability, which could mitigate the difficulties of genetic progress. This set of results reinforces the importance of multi-trait selection strategies in maize genetic improvement programs, optimizing gains in yield and associated traits.

Table 1 Genetic variability and descriptive parameters of grain yield (GY, kg ha^{-1}), prolificacy (PROL), seed mass (SM, g), days to anthesis (DA) and days to silking (DS), anthesis-silking interval (ASI), plant height (PH, m), and ear height (EH, m) in the PAB and PBA populations of full-sib maize progenies

Pop	Parameters	GY	PROL	SM	DA	DS	ASI	PH	EH
PAB	σ_g^2	1.37**	0.01**	4.98***	2.40	2.20***	0.21	0.01	0.01
	h_{gm}^2	0.52	0.48	0.75	0.82	0.82	0.36	0.58	0.72
	Accuracy	0.72	0.69	0.86	0.90	0.91	0.60	0.76	0.85
	Mean	9.37	1.22	32.30	67.81	68.98	1.17	2.54	1.50
	Minimum	7.14	1.09	28.80	65.02	66.07	0.76	2.34	1.29
	Maximum	11.15	1.47	37.26	70.63	72.46	1.82	2.67	1.67
PBA	σ_g^2	0.62	0.01**	4.06***	2.07	2.49***	0.26*	0.01	0.01
	h_{gm}^2	0.34	0.49	0.71	0.78	0.82	0.41	0.58	0.53
	Accuracy	0.58	0.70	0.84	0.88	0.91	0.64	0.76	0.73
	Mean	9.52	1.24	32.64	67.62	68.68	1.06	2.60	1.54
	Minimum	8.46	1.09	29.13	64.79	65.94	0.62	2.45	1.44
	Maximum	10.43	1.50	36.20	70.26	72.76	1.76	2.75	1.65

σ_g^2 = genetic variance; h_{gm}^2 = mean heritability. ***, **, * = significant at 0.001, 0.01, and 0.05, respectively, by the Likelihood Ratio Test (LRT), based on the chi-square (χ^2) distribution.

Observed correlations between the traits of the PAB and PBA populations guide selection and crossing strategies in maize breeding. In the PAB population, GY showed significant positive genetic and phenotypic correlations with PROL and SM. In contrast, a negative correlation was observed between GY and reproductive traits such as DA and ASI. In the PBA population, GY exhibited significant positive correlations with plant architecture traits PH and EH. Additionally, there were positive associations between PROL, PH, and EH, while SM showed a negative correlation with PROL but a positive correlation with EH and a negative correlation with ASI.

These relationships have direct implications for the improvement of each population. For the PA population, the positive correlations between GY, PROL, and SM suggest that selecting progenies with higher prolificacy and greater seed mass is an effective strategy to increase yield. This finding is supported by Shi et al. (2022), who emphasize seed mass as a crucial trait for field performance, and Faria et al. (2022) and Silveira et al. (2022), who highlight the number of ears per plant as an important trait for yield improvement. However, contrasting results were reported by Silva et al. (2023), who observed negative genetic correlations between prolificacy and yield in one of the studied populations, though they found positive correlations in another.

Furthermore, the negative correlation between GY and DA and ASI reflects a trend where plants with shorter reproductive cycles and more synchronized flowering exhibit better yield performance. This pattern aligns with the findings of Worku et al. (2016) and Benchikh-Lehocine et al. (2021), who indicate that reducing the cycle length and improving flowering synchrony are key factors for yield enhancement, especially under drought conditions. Recent studies, such as Almeida et al. (2024), suggest that controlling flowering time can be used to improve yield across different environmental conditions without compromising grain quality.

On the other hand, the PBA population showed a strong association between PH, EH, and GY, indicating that taller plants tend to be more yielding. This pattern, commonly observed in maize, was corroborated by Almeida et al. (2024), who linked plant height to greater nutrient assimilation and biomass, factors that contribute to higher grain production. However, this association should be interpreted cautiously, as increased height may raise the risk of lodging, especially in high-density plantings (Zhang et al. 2023). Thus, plant height must be balanced with other traits to ensure structural stability.

Additionally, the negative correlation between SM and PROL in the PBA population suggests that, as the number of ear per plant increases, the average seed mass tends to decrease. This can be explained by the allocation of plant resources, which need to be divided between the ear and the seeds. Similar results were obtained by Reichert Júnior et al. (2021). On the other hand, the positive correlation between SM and EH and the negative correlation with ASI, indicates that plants with higher seed mass tend to present higher EH and greater synchrony in flowering. However, this pattern diverges from Magar et al. (2021), who observed a positive relationship between SM and ASI.

Based on these observations, planning interpopulation crosses that combine desirable traits from the PAB and PBA populations is predicted to be an effective strategy for generating more yielding progenies. Integrating plants from the PAB population, which prioritize flowering synchrony and prolificacy, with plants from the PBA population, which exhibit greater height and yield associated with seed mass, could result in progenies with high yield potential and balanced architecture. As noted by Almeida et al. (2024), interpopulation crosses can also generate sufficient genetic variability to select lines that contribute to long-term sustainability.

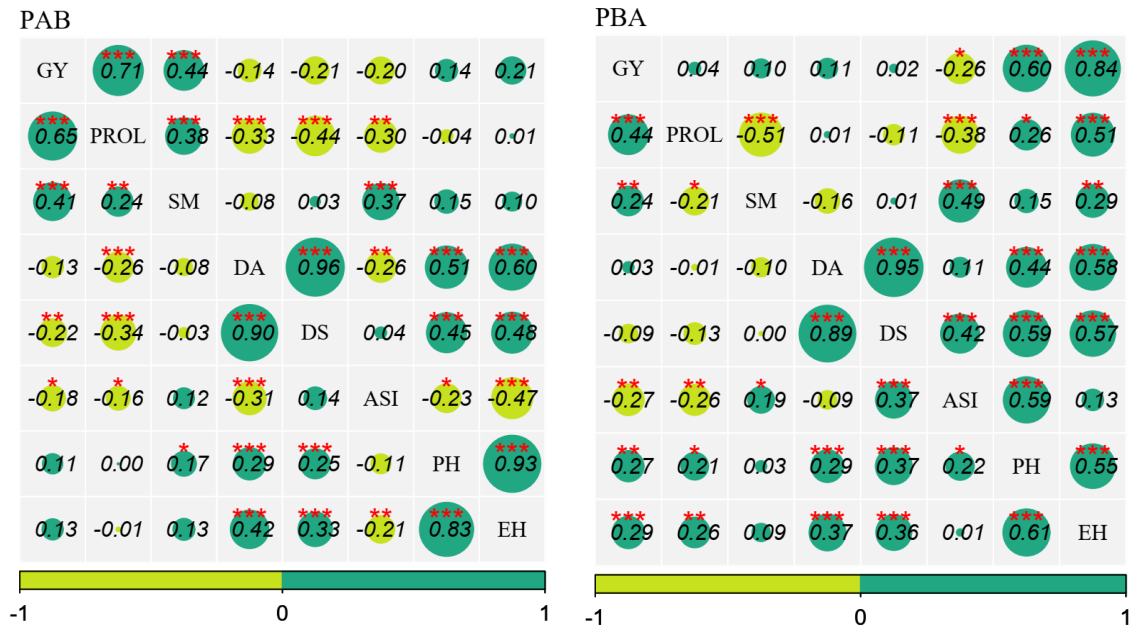


Fig. 1 Genetic correlations (above the diagonal) and phenotypic correlations (below the diagonal) among the traits grain yield (GY, kg ha-1), seed mass (SM, g), prolificacy (PROL), anthesis-silking interval (ASI), days to anthesis (DA), days to silking (DS), plant height (PH, m), and ear height (EH, m) in full-sib progenies of the PAB and PBA

maize populations. *, **, and *** indicate significance at 0.05, 0.01, and 0.001 probability levels, respectively, according to the t-test. The size of the circles is proportional to the magnitude of the correlation coefficients.

Results from regression analysis of trait subsets are presented, identifying the best trait combinations used in the construction of selection indices based on statistical criteria (Table 2). For the PAB and PBA populations, three and four top-performing models were selected, respectively, with the third and fourth models being full models (FM) that included all traits. In the PAB population, the M2 model outperformed the others, exhibiting the best Bayesian Information Criterion (BIC) and Mallows' Cp values, followed by the FM, both demonstrating performance fit. Conversely, in the PBA population, the M3 and FM models provided the best fit, showing the lowest BIC and Cp values. The traits corresponding to each model are detailed in Tables 3 and 4.

Table 2 Adjusted coefficient of determination (R^2 adj), Bayesian information criterion (BIC), and Mallows' Cp value of the best models (M) and the full model (FM) for trait combinations of the PAB and PBA populations of full-sib maize progenies.

Pop	Model	R^2_{adj}	BIC	Cp
PAB	M1	0.46	-94.10	48.00
	M2	0.58	-130.00	3.77
	FM	0.58	-127.00	4.00
PBA	M1	0.33	-45.90	73.70
	M2	0.51	-86.10	17.80
	M3	0.55	-93.20	7.30
	FM	0.56	-92.70	5.00

Table 3 presents the results of the selection response for full-sib progeny in the maize PAB, evaluating different trait combinations and selection index strategies (FAI-BLUP, MGIDI, and SH). The objective was to investigate the impact of these combinations on the selection of high-performance progeny.

Initially, in the PAB, M1 considered the traits GY and PROL, resulting in consistent gains for both traits across all selection index strategies, with SH standing out with the highest SSE (99.92%). Next, M2, which included GY, PROL, and SM, showed responses similar to M1 for GY and PROL. However, the SH index exhibited a reduction in selection gain for GY and PROL, accompanied by an increase in SM.

In the FM model, the trait DS was added. In these combinations, there was a reduction in GY gain across all selection index strategies. On the other hand, DIS demonstrated the highest gain for GY, proving to be a robust approach for maximizing grain yield without simultaneously considering other traits. Meanwhile, the M2 model combined with FAI-BLUP or MGIDI provided simultaneous trait gains for maize improvement, with a desirable response for GY achieving 99.33% SSE, close to DIS, and 1.01% SSE over FM.

Table 3 Selection response of the best models (M) of combinations of traits for simultaneous selection indices factor analysis and ideotype design (FAI-BLUP), Multitrait Genotype-Ideotype Distance (MGIDI) and Smith and hazel (SH) via best linear unbiased prediction in full-sib maize progenies of the PAB population.

Model	FA	Traits	FAI_BLUP		MGIDI		SH		Ideotype	Sense
			XS	SG%	XS	SG%	XS	SG%		
M1	1	GY	10.41	11.10	10.41	11.10	10.47	11.75	Max	Increase
	1	PROL	1.33	9.32	1.33	9.32	1.33	8.64	Max	Increase
M2	1	GY	10.41	11.10	10.41	11.10	9.90	5.74	Max	Increase
	1	PROL	1.33	9.32	1.33	9.32	1.25	2.37	Max	Increase
	1	SM	33.10	2.45	33.10	2.45	35.31	9.31	Max	Increase
FM	1	GY	10.30	9.99	10.20	8.91	9.46	1.05	Max	Increase
	1	PROL	1.33	8.63	1.33	9.26	1.23	0.85	Max	Increase
	2	SM	33.40	3.34	33.40	3.52	34.99	8.32	Max	Increase
	2	DS	68.20	-1.13	67.90	-1.50	69.53	0.79	Min	Decrease
DIS		GY	10.48	11.91					Max	Increase

Selected model with an SSE of 99.33 near DIS and 1.01 over FM: M2 + MGIDI.

SSE % = Simultaneous selection efficiency close to direct selection (DIS) and over the full model (FM) for grain yield (GY) of the best combination model of traits and simultaneous selection index. XS = population mean, SG = Selection gain. The factors (FA), selection directions and ideotypes are terms applied only to the FAI-BLUP or MGIDI indices.

In the analysis of the PBA population (Table 4), models M1, M2, and M3 stood out as those delivering the highest genetic gains in GY under the SH, MGIDI, and FAI-BLUP selection strategies, respectively. The combination of M3 with the MGIDI index maximized genetic gains in GY over FM and near DIS (SSE% = 1.00 and 98.92, respectively), while also driving desirable gains in associated traits (PROL, SM, and ASI). This subset of traits exhibits genetic parameters of favorable magnitude, underscoring their importance in maintaining genetic variability and developing high-performance progenies in the reciprocal selection cycle.

The PROL trait is widely recognized as a relevant criterion for improving yield (Faria et al., 2022; Silveira et al., 2022). However, studies such as Almeida et al. (2024) report a reduction in GY gains when predicted by PROL, particularly in cases of negative correlation between these traits. In this study, the prediction of GY based on PROL was undesirable in FM, and gains in PROL were unfavorable across all models associated with SH. This behavior can be attributed to a zero genetic correlation with GY, as well as negative phenotypic and genetic correlations with SM, this trait showing the highest response in this population and a negative genetic correlation with ASI, which exhibited the largest undesirable genetic gain.

These results differed from those observed in all models applied to the FAI-BLUP and MGIDI indices. The divergence in results is partly because the SH index is highly dependent on genetic and phenotypic covariance matrices, as well as economic values, according to Smith (1936) and Hazel (1943). This may limit the selection for traits with low genetic correlation. In contrast, non-parametric methods such as MGIDI and FAI-BLUP prioritize a balance among multiple traits, regardless of genetic correlations, allowing simultaneous gains even in scenarios with low genetic relationships between evaluated traits (Rocha et al. 2018; Olivoto et al. 2019).

Additionally, Worku et al. (2016) and Benchikh-Lehocine et al. (2021) highlight that ASI is an indicator of stress tolerance and pollination efficiency, with lower values being desirable for increased yield. In the models associated with FAI-BLUP and MGIDI strategies, a reduction in ASI was observed, indicating the effectiveness of these strategies in selecting plants with better reproductive synchrony. However, the SH index showed an increase in ASI in the M3 and FM models, suggesting a lower selection capacity for this trait in PBA. On the other hand, the SH index proved particularly useful for higher gains in SM across models M1 to M3, suggesting an effective strategy for improving seed mass. The presented results demonstrate that different subsets of traits, combined with simultaneous selection indices, influence selection efficiency for multiple traits in maize progenies. Overall, the MGIDI simultaneous selection strategy based on the M3 model proved efficient in PBA.

Table 4. Selection response of the best models (M) for trait combination in simultaneous selection indices: factor analysis and ideotype design via best linear unbiased prediction (FAI-BLUP), Multitrait Genotype-Ideotype Distance (MGIDI) and Smith and Hazel in full-sib maize progenies from the PBA population.

Model	FA	Traits	FAI_BLUP		MGIDI		SH		Ideotype	Sense
			XS	SG%	XS	SG%	XS	SG%		
M1	1	GY	10.03	5.43	10.03	5.43	10.09	6.07	Max	Increase
	1	PROL	1.34	8.25	1.34	8.25	1.24	-0.03	Max	Increase
M2	1	GY	10.10	6.10	10.09	5.99	9.62	1.05	Max	Increase
	1	PROL	1.26	1.62	1.32	7.02	1.20	-2.83	Max	Increase
	1	SM	33.9	3.87	33.2	1.82	34.85	6.78	Max	Increase
M3	1	GY	10.09	5.91	10.14	6.58	9.55	0.39	Max	Increase
	1	PROL	1.27	2.86	1.30	5.22	1.20	-2.72	Max	Increase
	2	SM	33.8	3.63	33.3	2.10	34.83	6.71	Max	Increase
	1	ASI	1.06	-1.56	0.89	-15.6	1.24	17.00	Min	Decrease
FM	1	GY	9.97	4.80	10.10	5.76	9.47	-0.45	Max	Increase
	1	PROL	1.27	2.64	1.31	5.78	1.20	-2.64	Max	Increase
	1	SM	34.10	4.39	33.6	2.95	33.93	4.04	Max	Increase
	2	DS	68.20	-0.72	68.40	-0.46	70.28	2.32	Min	Decrease
	2	ASI	1.04	-1.67	0.93	-12.20	1.29	21.62	Min	Decrease
DIS	_	GY	10.20	6.82	_	_	_	_	Max	Increase

Selected model with an SSE of 98.92 near DIS and 1.00 over FM: M3 + MGIDI.

SSE % = Simultaneous selection efficiency close to direct selection (DS) and over the full model (FM) for grain yield (GY) of the best combination model of traits and simultaneous selection index. XS = population mean, SG = Selection gain. The factors (FA), selection directions and ideotypes are terms applied only to the FAI-BLUP or MGIDI indices. However, the selection directions of these indices were adapted to the SH.

Fig. 2 presents the selected high-performance progenies for the PAB and PBA populations using simultaneous selection index strategies based on higher SSE% and balanced traits. For the PAB population, the MGIDI index was selected based on the subset of traits from model M2. For the PBA population, the MGIDI index associated with the subset of traits from model M3 was selected.

Among the top 20% of selected progenies from PAB and PBA, 69.20% and 80.00%, respectively, were common between simultaneous and direct selection (Fig. 3). The high-performance progenies identified hold significant agronomic value for the breeding program, combining desirable traits such as high yield potential, prolificacy, short anthesis-silking interval, and seed mass. These attributes promote a more targeted selection response aligned with the program's objectives. In the context of maize breeding programs, integrating trait combination models and simultaneous selection indices serves as an optimal tool for maximizing genetic gain, enabling more precise selection of high-performance multi-trait progenies for future evaluations. Previous studies, such as those by Rocha et al. (2018) and Olivoto and Nardino (2020), have already demonstrated that using selection indices is an efficient strategy for optimizing selection response by integrating multiple traits into a single metric.

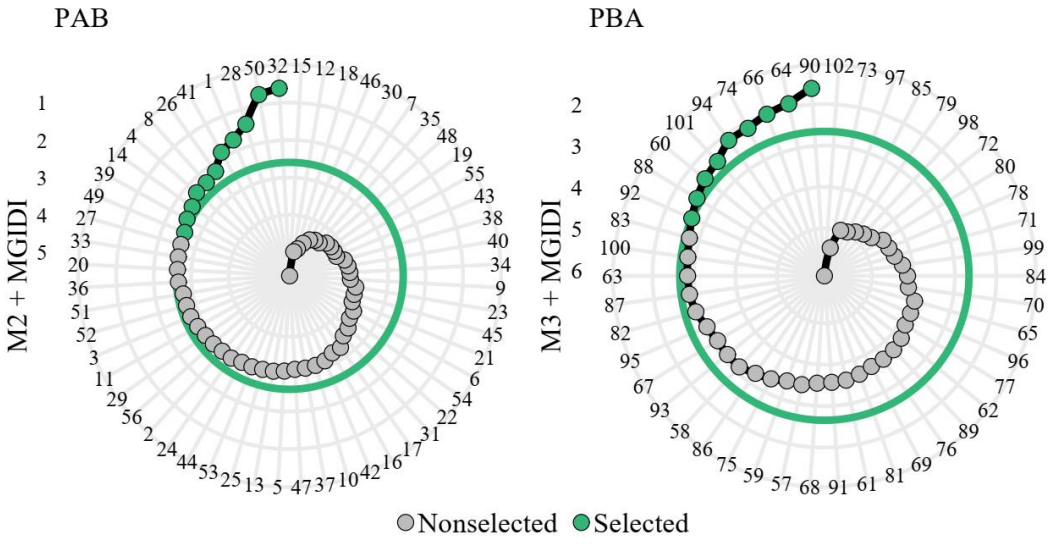


Fig. 2 Maize progenies selected by the best models of combinations of traits and simultaneous selection indices in the PAB and PBA populations.

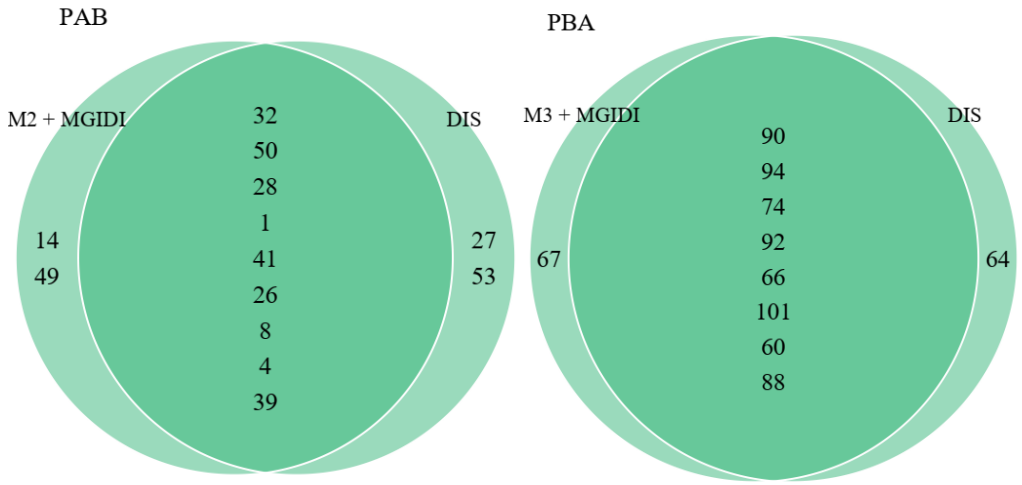


Fig. 3 Common progenies in the intersection region of the diagrams, selected by the best models (M2 and M3) of trait combinations and simultaneous selection indices (MGIDI) and direct selection (DIS) in PAB and PBA maize populations.

Conclusion

The M2 + MGIDI and M3 + MGIDI models of the GY+PROL+SM and GY+PROL+SM+ASI trait combinations provide higher simultaneous selection efficiency (SSE%) of 99.33 and 98.92 for PAB and PBA close to direct selection with identification of high-performance multi-trait progenies and outperformed the use of all traits by the full model with SSE% of 1.01 and 1.00, respectively.

The study demonstrated that the integration of the Best Subset Regression (BSR) methodology with Simultaneous Selection Indices (SSI) is an effective strategy to maximize genetic gains by selecting high-performance multi-trait progenies in reciprocal recurrent selection of maize.

Future studies can apply, verify and validate these selection strategies in different experimental contexts.

Acknowledgements

To Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, finance code 001), Brazil, for supporting this research.

Statements & Declarations

"This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, finance code 001), Brazil, for supporting this research."

"The authors declare that no funds, grants, or other support were received during the preparation of this manuscript."

Data availability The data that support the findings of this study are available from the corresponding author CP upon reasonable request.

Conflict of interest The authors declare no competing interests

Reference

Almeida PHS, Vilela VJB, Torres IY, et al (2024) Genetic potential of maize populations derived from commercial hybrids for interpopulation breeding. Rev Caatinga 37:1–7. https://doi.org/10.1590/1983-21252024v3711736rc

Benchikh-Lehocine M, Revilla P, Malvar RA, Djemel A (2021) Response to selection for reduced anthesis-silking interval in four algerian maize populations. Agronomy 11:1–11. https://doi.org/10.3390/agronomy11020382

BORÉM, A; GALVÃO, J. C. C.; PIMENTEL MA (2017) No TitleMilho: do plantio à colheita., 2nd edn. Viçosa

Brooks GP, Ruengvirayudh P (2016) Best-Subset Selection Criteria for Multiple Linear Regression. Gen Linear

Model J 42:14–25

Costa M, Reis DOS (2009) Viabilidade da seleção recorrente recíproca em populações derivadas de híbrido simples de milho

Cruz, Cosme Damião; Regazzi, Afair Jose; Carneiro PC souza (2012) Modelos biométricos aplicados ao melhoramento genético, 4th edn. Vicosa

de Faria SV, Zuffo LT, Rezende WM, et al (2022) Phenotypic and molecular characterization of a set of tropical maize inbred lines from a public breeding program in Brazil. BMC Genomics 23:1–17. https://doi.org/10.1186/s12864-021-08127-7

de Resende MDV (2016) Software Selegen-REML/BLUP: A useful tool for plant breeding. Crop Breed Appl Biotechnol 16:330–339. https://doi.org/10.1590/1984-70332016v16n4a49

eugenics HS-A of, 1936 undefined (1936) A discriminant function for plant selection Cesar. Wiley Online Libr 7:240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

Genetics LH-, 1943 undefined The genetic basis for constructing selection indexes Cesar. academic.oup.com

Humberto Gonçalves dos Santos, Paulo Klinger Tito Jacomine, Lúcia Helena Cunha dos Anjos, Virlei Álvaro de Oliveira, José Francisco Lumbreras, Maurício Rizzato Coelho, Jaime Antonio de Almeida, José Coelho de Araújo Filho, João Bertoldo de Oliveira TJFC (2018) No TitleSistema Brasileiro de Classificação de Solos, 5th edn. Empresa Brasileira de Pesquisa Agropecuária Embrapa Solos, Embrapa Brasília, DF

Li H, Liu K, Li Z, et al (2023) Mixing trait-based corn (Zea mays L.) cultivars increases yield through pollination synchronization and increased cross-fertilization. Crop J 11:291–300. https://doi.org/10.1016/j.cj.2022.05.007

Magar BT, Acharya S, Gyawali B, et al (2021) Genetic variability and trait association in maize (Zea mays L.) varieties for growth and yield traits. Heliyon 7:e07939. https://doi.org/10.1016/j.heliyon.2021.e07939

Olivoto T, Lúcio ADC (2020) metan: An R package for multi-environment trial analysis. Methods Ecol Evol 11:783–789. https://doi.org/10.1111/2041-210X.13384

Olivoto T, Lúcio ADC, da Silva JAG, et al (2019) Mean performance and stability in multi-environment trials II: Selection based on multiple traits. Agron J 111:2961–2969. https://doi.org/10.2134/AGRONJ2019.03.0221

Olivoto T, Nardino M (2021) Genetics and population analysis MGIDI : toward an effective multivariate selection in biological experiments. 37:1383–1389. https://doi.org/10.1093/bioinformatics/btaa981

Paul PA, Munkvold GP (2005) Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. Phytopathology 95:388–396. https://doi.org/10.1094/PHYTO-95-0388

Reichert Júnior FW, Ogliari JB, Maghelly OR, Souza R de (2021) Relationship between phenological and morphological characteristics of plant with popcorn races of a diversity microcenter in southern Brazil. Res Soc Dev 10:e46710716734. https://doi.org/10.33448/rsd-v10i7.16734

Rocha JR do AS de C, Machado JC, Carneiro PCS (2018) Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy. GCB Bioenergy 10:52–60. https://doi.org/10.1111/gcbb.12443

Shi D, Hang J, Neufeld J, et al (2022) Estimation of crude protein and amino acid contents in whole, ground and defatted ground soybeans by different types of near-infrared (NIR) reflectance spectroscopy. J Food Compos Anal 111:104601. https://doi.org/10.1016/j.jfca.2022.104601

Silva, É. M. D., Guedes, M. L., Crispim Filho, A. J., Ciappina, A. L., Reis, E. F. D., & Resende MPM (2023) Genetic parameters and selection for multiple traits in recurrent selection populations of maize 1. Rev Ceres 70:81–90

Silveira ES, Silva AF, Santos BN, et al (2022) Morphological characterization and selection of maize genotypes

for the semiarid region. 1–17

Worku M, Makumbi D, Beyene Y, et al (2016) Grain yield performance and flowering synchrony of CIMMYT's tropical maize (Zea mays L.) parental inbred lines and single crosses. Euphytica 211:395–409. https://doi.org/10.1007/s10681-016-1758-3

Zhang P, Gu S, Wang Y, et al (2023) The relationships between maize (Zea mays L.) lodging resistance and yield formation depend on dry matter allocation to ear and stem. Crop J 11:258–268. https://doi.org/10.1016/j.cj.2022.04.020

ARTIGO 2- variance components and factors analysis based on phenotypic and genetic correlations among bulk-pollinated maize populations

Periódico: Crop improvement, versão preliminar.

César Pedro (https://orcid.org/0000-0002-2963-8652)[a] and João Cândido de Souza (https://orcid.org/0000-0001-9580-4631)[a]*

[a]*Departamento de Biologia, Universidade Federal de Lavras, Lavras, Brasil*

*e-mail: cansouza@ufla.br

Abstract: This study aimed to evaluate variance components and perform factor analysis based on phenotypic and genotypic correlations to select intracross- and intercross-populations maize progenies. Using an alpha lattice design, populations derived from intracross (PA and PB) and intercross (PAB and PBA) crosses by Bulk Pollen Pollination (BPP) were evaluated. Variances and means were estimated via restricted maximum likelihood (REML) and Best Linear Unbiased Prediction (BLUP), respectively. BLUP values supported correlations used in factor analysis and scores in k-means clustering analysis. High genetic potential was observed in PB, with the highest additive variance for grain yield: GY (16.50) and anthesis-silking interval: ASI (18.51), along with notable values for seed mass: SM (34.27) and days to female flowering: DFF (19.34). Among inter-population crosses, PAB exhibited the highest additive variance for SM (59.46) and DFF (25.84), while PBA stood out in GY (15.28), all associated with high narrow-sense heritability. The analysis revealed genetic and phenotypic variability, indicating complementarity between populations PA and PB in

generating PAB and PBA hybrids. Factorial selection strategies in BPP populations proved effective in identifying progenies with simultaneous phenotypic and genetic performance, contributing to the development of cultivars with greater agronomic and genetic stability.

Keywords: Zea mays L., additive variance, narrow-sense heritability, k-means.

Variance components and factors analysis based on phenotypic and genetic correlations among Bulk-pollinated corn populations

Resumo: Objectivou-se avaliar componentes de variância e realizar análise fatorial baseada em correlações fenotípicas e genotípicas para selecionar progênies de milho intra e interpopulacionais. Utilizando delineamento alpha lattice, populações derivadas de cruzamentos intra- (PA e PB) e interpopulacionais (PAB e PBA) por Polinização em Massa foram avaliadas. As variancias e as medias foram estimadas via máxima verossimilhança restrita e BLUP, respectivmente. Os valores BLUP subsidiaram correlações usadas na análise fatorial e os escores na análise de agrupamento por k-means. Observou-se alto potencial genético em PB, com maior variância aditiva para GY (16,50) e ASI (18,51), além de valores notáveis para SM (34,27) e DFF (19,34). Entre as interpopulacionais, PAB apresentou a maior variância aditiva para SM (59,46) e DFF (25,84), enquanto PBA destacou-se em GY (15,28), todas associadas a herdabilidades restritas elevadas. A análise revelou variabilidade genética e fenotípica, apontando complementaridade entre populações A e B na geração dos híbridos PAB e PBA. As estratégias de seleção fatorial em populações BPP mostraram eficácia ao identificar progênies com desempenho simultâneo em relacao a expressão fenotípica e genética, contribuindo para desenvolvimento cultivares de maior estabilidade agronômica e genetica.

Palavras-chave: Zea mays L., variância aditiva, herdabilidade restrita, k-means.

Introduction

Understanding the magnitude of variance components is crucial for unraveling genetic variability in maize populations under a reciprocal recurrent selection (RRS) scheme. In this context, accurate estimates of additive variances, narrow-sense heritability, and phenotypic and genetic correlations among traits are essential for differentiating the genetic and agronomic potential of maize populations, guiding selection and breeding strategies.

Additive variance plays a key role, as it represents the portion of genetic variability directly transmitted from one generation to the next (Cruz et al. 2014) and is one of the main determinants of selection gains (Silva-Díaz et al. 2018). Narrow-sense heritability reflects the proportion of phenotypic variation attributed to additive variance, providing a measure of the predictability of selection success (Ramalho et al. 2012).

The selection of high-performance genotypes in plant breeding programs can be optimized by analyzing genetic and phenotypic correlations among traits. While phenotypic correlation reflects observable associations influenced by genetic and environmental effects (Falconer 1996, Cabral 2011), genetic correlation estimates the heritable relationship between traits resulting from additive genetic effects, enabling more precise selection (Falconer, 1996).

Among crossing methods used to quantify genetic variability in maize populations, Bulk Pollen Pollination (BPP) has emerged as an efficient strategy (Wang et al. 2019). This methodology, applied in maize genetic crosses, promotes the mixing and recombination of pollen from different genotypes, enhancing population genetic diversity (Talabi et al. 2017). In the intra- and inter-population crosses proposed in this study, BPP could expand the genetic potential of populations and support the reciprocal recurrent selection (RRS) program for maize half-sib progenies.

However, the success of RRS depends not only on the available genetic variability but also on the accurate identification and ranking of high-performance mult-trait progenies. Various quantitative genetics based methodologies, such as factor analysis (Rocha et al. 2018; Olivoto and Nardino 2021), have successfully assisted in selecting high-performance multi-trait genotypes. However, these analyses are typically based on phenotypic correlations for progeny ranking. In this study, we propose a factor analysis based on phenotypic and genetic correlation

data of half-sib progeny traits derived from intra- and inter-population crosses using the BPP method.

Thus, this study aims to estimate additive variance, narrow-sense heritability, and phenotypic and genetic correlations among traits, and to select high-performance progenies through factor analysis based on simultaneous phenotypic and genetic expression. With this strategy, we expect to identify progenies that can replicate their performance in future breeding generations, increasing selection reliability and the efficiency of genetic gains in the SRR program.

Material and methods

Two experiments were conducted in the experimental area of Muquém Farm (21°12'07.2"S, 44°58'44.3"W, at an altitude of 900 m), which belongs to the Center for Scientific and Technological Development in Agriculture at the Federal University of Lavras. The farm is located in the southern region of Minas Gerais, Brazil. The soil in the experimental area is classified as Oxisol (Latossolo), according to the classification by **Santos et al. (2018).** The region has a humid temperate climate (Cwa), characterized by dry winters and rainy summers.

The first experiment was conducted between November 2021 and March 2022, with an average temperature of 23.74°C and rainfall of 232.56 mm. The second experiment took place between March and July 2024, with an average temperature of 22.96°C and cumulative rainfall of 218.4 mm during the experimental period, as recorded by the Lavras weather station (code: 83687, geographic coordinates: 21°13'34.0"S, 44°58'47.0"W).

First evaluation, recombination, and Bulk Pollen Pollination in Maize Populations

A total of 102 full-sib maize progenies were evaluated, together with two double-crossed commercial hybrids (checks), using an alpha-lattice design with three replicates and 24 blocks. A detailed description of the experimental procedures is provided in Pedro et al. (2023). Among the 102 progenies, 15% were selected based on genetic divergence and agronomic performance (Pedro et al., 2022). In November 2022, S1 progenies from the PA and PB populations corresponding to the selected lines were sown and subjected to intragroup

recombination in a full diallel crossing scheme. The resulting progenies were mass harvested, thus reconstituting the PA and PB populations.

From October 2023 to March 2024, a total of 4,000 plants (2,000 from each population, PA and PB) were planted at a spacing of 60 × 25 cm in two 20 × 15 m blocks, separated by a two-meter buffer. From these populations, four half-sib progeny populations (intra- and inter-population) were developed using the Bulk Pollen Pollination (BPP) method as described by Wang et al. (2019).

To generate the intra-population progenies (PA and PB), pollen from 10 randomly selected plants was collected, mixed, and used to pollinate the lower ear of a plant from the same population. The same procedure was followed for population PB. For the inter-population progenies, pollen from population PA was used to pollinate plants from population PB, and vice versa (Figure 1).

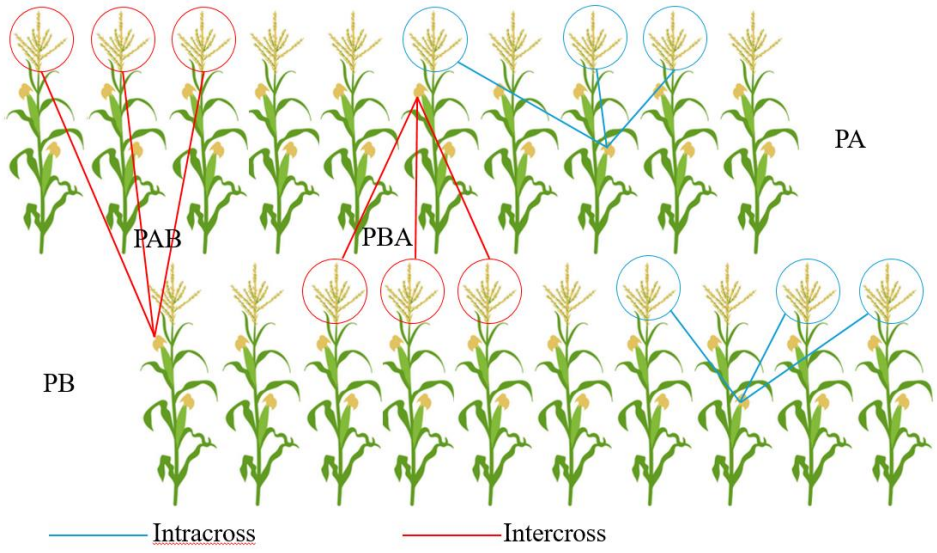


Figure 1: Breeding scheme for intracross populations (PA and PB) and intercross populations (PAB and PBA) using the Bulk Pollen Pollination method.

Second evaluation and selection of maize populations

For each population, 78 progenies were selected based on good appearance, well-filled ears, and sufficient kernel count in both ears, totaling 312 progenies plus two commercial double-cross hybrids (checks: Check1-Ufla JM 100, Check2-Pioneer 3707).

The trial was evaluated in a 10 × 8 alpha-lattice design with two replications. Sowing took place between March 5 and 6, 2024, at a density of four seeds per linear meter in 4 m-long plots spaced 0.6 m apart. Basal fertilization consisted of 250 kg ha^{-1} of NPK fertilizer (8% N, 28% P_2O_5, and 16% K_2O). A topdressing of 200 kg ha^{-1} urea-N (45% N) was applied 25 days after sowing. Crop management followed region-specific recommendations (Borém 2017)

Evaluated traits and data analysis

The evaluated traits and results of the first experiment can be found here (Pedro et al. 2023). In the second experiment, the following traits were assessed: grain yield (GY, t ha^{-1}), seed mass (SM, g), days to female flowering (DFF), and anthesis-silking interval (ASI).

Variance components, including phenotypic variance ($\hat{\sigma}_p^2$), genetic variance ($\hat{\sigma}_g^2$), and environmental variance ($\hat{\sigma}_e^2$), were estimated using REML via the lme4 package (Bates et al. 2015). Additionally, additive variance ($\hat{\sigma}_A^2$) and narrow-sense heritability (h_r^2) were calculated for half-sib families following the method of Comstock and Robinson (1952) as follows:

$$\hat{\sigma}_A^2 = 4\sigma_{hs}^2$$

Where: σ_{hs}^2 = Variance component due to genetic differences among half-sib families.

$$h_r^2 = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$$

According to Johnson et al. (1955), heritability estimates were classified as high (>60%), moderate (30–60%) and low (<30%). Phenotypic and genetic correlation analyses were performed following Cruz et al. (2014), followed by Bartlett's sphericity test (Bartlett, 1951) using the EFAtools package (Markus et al. 2024). to assess the significance of correlation matrices relative to the identity matrix. Subsequently, eigenvalues were obtained for factor extraction with varimax rotation, retaining the first two factors with the highest explained variance, followed by the derivation of progeny scores.

The factor scores obtained from phenotypic and genotypic correlations were standardized, and the relationship among progenies and their traits was visualized in a 2D system based on the two most variable factors. Progenies were classified into four quadrants. K-means clustering (Hartigan, 1979) was applied using the factoextra package (Kassambara & Mundt, 2020), with k = 4 determined by the Elbow method, to reduce ambiguity in selecting homogeneous progenies within their respective quadrants. Finally, a coincidence analysis was

conducted for the selected progenies based on phenotypic and genotypic correlation factors to enable simultaneous selection of progenies with both phenotypic and genetic expression.

Results and Discussion

Table 1 presents variance components to unravel the genetic and phenotypic variability in intrapopulation (PA and PB) and interpopulation (PAB and PBA) maize crosses developed using the BPP methodology. In PA, genetic variance was significant for all traits except ASI. Additive variance ranged from 0.63 (ASI) to 24.7 (SM), while narrow-sense heritability varied from 0.32 (ASI) to 0.81 (GY), with selection accuracy between 0.43 (ASI) and 0.82 (GY). These parameters were most pronounced for SM, DFF, and GY, with SM exhibiting the highest additive variance among the population's traits.

In the PB population, significant genetic variance was observed for all traits. Narrow-sense Heritability was high for GY (0.93) and SM (0.85), with high selection accuracy across all traits, reaching 0.94 for GY. The additive genetic variance for GY (16.50) was significantly higher than in the other populations, and for SM (34.27), it was greater than in the PA and PBA populations.

For the interpopulation PAB, genetic variance was significant for all traits except ASI. Narrow-sense heritability was high for GY (0.86) and DFF (0.85), suggesting strong genetic influence on these traits. Additive variance was notably high for SM (59.46), surpassing all other populations, while GY (8.55) exceeded that of PA. In PBA, significant genetic variance was detected for all traits. Narrow-sense heritability was moderate for SM (0.57). Additive variance was moderate for SM (12.79), high for GY (15.28) with the lowest values for DFF (4.71) and ASI (1.62).

Table 1. Phenotypic ($\hat{\sigma}_p^2$), genetic ($\hat{\sigma}_g^2$), and additive ($\hat{\sigma}_A^2$) variance components, narrow-sense heritability (h_r^2), and selection accuracy for grain yield (GY, kg ha^{-1}), seed mass (SM, g), days to female flowering (DFF), and anthesis-silking interval (ASI) in intrapopulations (PA, PB) and interpopulations (PAB, PBA) maize populations.

Traits	$\hat{\sigma}_g^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_A^2$	h_r^2	Accuracy	Mean	Min	Max
	Intracross: PA							
GY	1.38 ***	2.71	5.51	0.81	0.82	4998	3342	10009
SM	6.02 ***	15	24.07	0.75	0.76	28.57	22.77	33.41
DFF	1.56 ***	3.47	6.24	0.77	0.81	63.77	60.62	67.52

ASI	0.16	1.68	0.63	0.29	0.43	-0.23	-1.48	0.85
				Intracross: PB				
GY	4.12 ***	5.3	16.5	0.93	0.94	5345	1697	10510
SM	8.57 ***	14.8	34.27	0.85	0.86	25.81	20.24	31.22
DFF	4.83 ***	9.23	19.34	0.82	0.83	67.02	63.47	70.51
ASI	4.63 **	15.4	18.51	0.63	0.68	0.07	-2.90	3.2
				Intercross: PAB				
GY	2.14 ***	3.79	8.55	0.86	0.86	6504	3963	11279
SM	14.90 **	32.6	59.46	0.78	0.79	34.4	27.96	41.98
DFF	6.46 **	8.85	25.84	0.92	0.92	63.89	59.26	72.06
ASI	0.30	1.68	1.20	0.47	0.55	-0.40	-1.98	0.80
				Intercross: PBA				
GY	3.82 ***	9.13	15.28	0.74	0.77	6960	3957	10864
SM	3.20 *	13.2	12.79	0.57	0.63	29.25	25.82	31.62
DFF	1.18 ***	3.52	4.71	0.70	0.72	63.08	60.72	66.56
ASI	0.40 **	1.29	1.62	0.65	0.70	0.11	-0.66	2.64

The findings reveal significant differences in additive genetic variance and narrow-sense heritability between populations derived from intra- and inter-population crosses using the BPP method. While Wang et al. (2018) demonstrated the effect of BPP on genome representativeness in maize populations, this study shows that the method also effectively enhanced the mean and additive genetic variability in PAB and PBA populations, a key insight for breeding programs aiming to improve reciprocal recurrent selection potential.

The PAB population exhibited the highest expression of additive variance for grain yield, seed mass, and days to female flowering, with PB as the parental source contributing most to this variability. Additionally, PAB demonstrated a more balanced distribution of additive variance across all traits. According to Pedro et al. (2023) and Almeida et al. (2024), populations with greater additive variance and heritability tend to respond better to selection. In contrast, the low and moderate heritability observed for ASI in the PA and PAB populations, respectively, is due to non-significant genetic variability and suggests that this trait is more influenced by environmental factors, as discussed by Li et al. (2023). This necessitates alternative approaches to improve selection efficiency.

Table 2 presents the fitness parameters of the factor analysis based on genetic and phenotypic correlation matrices of maize population traits. The chi-square values indicate that Bartlett’s sphericity test was significant for all populations, both for phenotypic and genetic

correlations. This confirms that the correlation matrices significantly deviate from the identity matrix (Bartlett, 1951), validating the suitability of factor analysis.

Table 2 Bartlett's sphericity test for the suitability of factor analysis (FA) based on phenotypic (rf) and genetic (rg) correlations of traits in half-sib maize progenies, derived from intracross (PA and PB) and intercross (PAB and PBA) populations using the BPP method.

Cross: Populations	Type of correlations	X^2	*p.value*	Eigenvalues	
				FA1	FA2
Intarcross: PA	Rf	18.31	0.006	1.47	1.03
	Rg	70.56	0.001	1.74	1.43
Intracross: PB	Rf	71.19	0.0001	2.07	0.96
	Rg	189.49	0.0001	2.44	0.85
Intercross: PAB	Rf	15.64	0.0158	1.53	1.05
	Rg	24.27	0.0005	1.70	0.98
Intercross: PBA	Rf	58.10	0.0001	1.86	1.01
	Rg	144.82	0.0001	2.12	1.22

X^2 = chi-square test.

The magnitudes of the eigenvalues suggest that most of the variance is explained by FA1 in all populations, indicating underlying key traits that control phenotypic and genetic variation among progenies within and between populations.

Figure 1 displays the distribution of maize progenies from populations PA, PB, PAB, and PBA based on FA1 and FA2 factors, considering yield-related traits (GY and SM) and flowering traits (ASI and DFF). The reciprocal parental populations, PA and PB, exhibited distinct patterns of variability and associations between yield and flowering traits. The derived hybrid populations, PAB and PBA, reflect the combination of traits inherited from the parental populations. In population PA, phenotypic analysis revealed that FA1 (accounting for 60.05% of variability) was strongly correlated with yield, while FA2 (29.16%) was associated with flowering, highlighting progenies from Groups 2 and 3. Genotypic confirmation maintained this pattern (FA1 = 55.29%; F2 = 40.01%), with superior progenies in Groups 1 and 2.

In contrast, population PB displayed a different trend: FA1 (85.29%) was predominantly linked to flowering, while FA2 (13.16%) was related to yield, identifying promising progenies in Groups 2 and 4. At the genotypic level, results were similar (FA1 = 89.33%; FA2 = 10.61%), with notable performance in Groups 1 and 4.

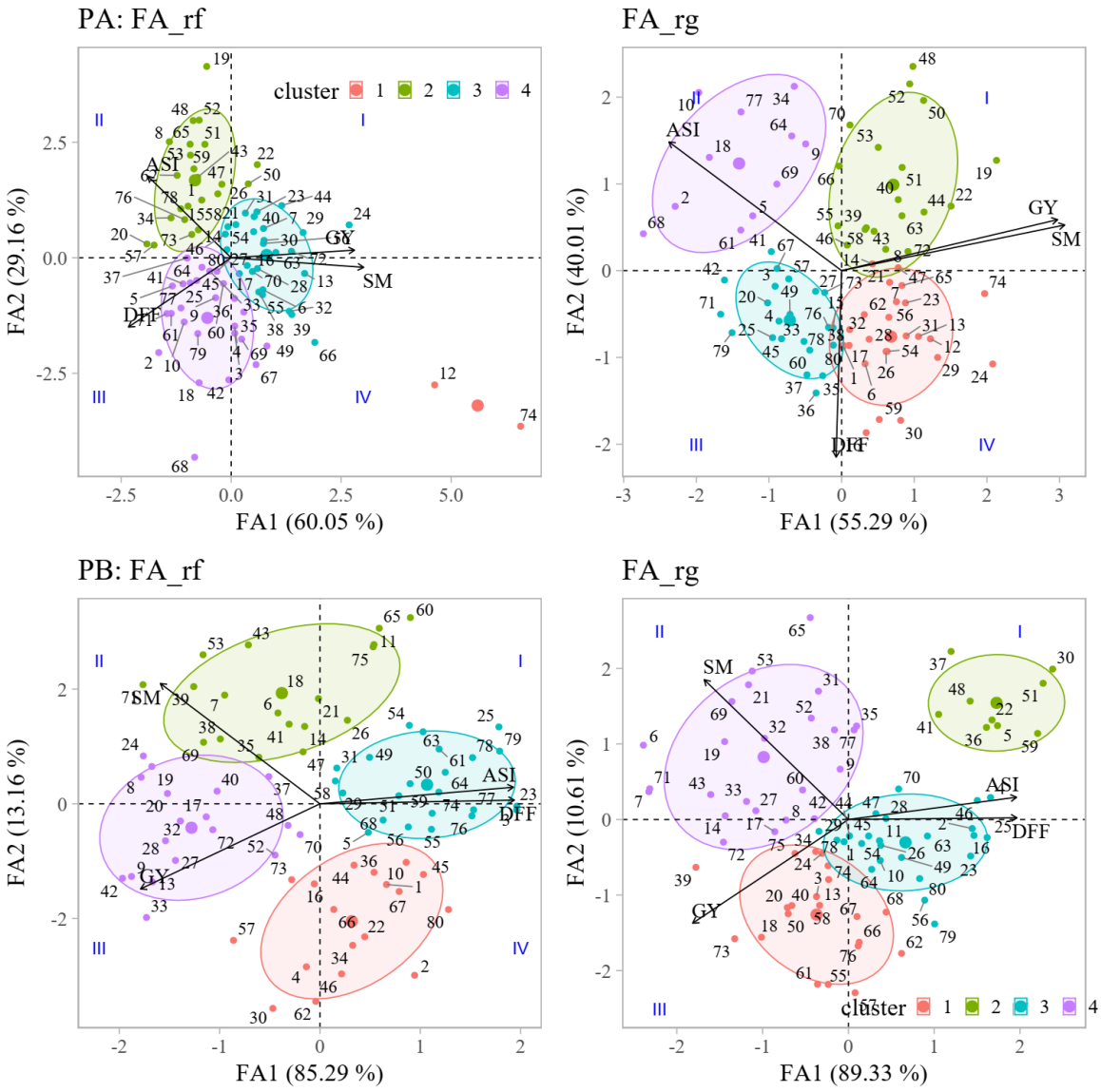


Figure 2. Factor analysis based on phenotypic (rf) and genotypic (rg) correlations in maize progenies from PA and PB intracross populations. The graphs display the projection of the progenies in the two-dimensional plane formed by the two factors (Factor), with distinct clusters (groups) identified by colors. Each progeny is represented by numbers, and the ellipses outline the clusters based on grain yield (GY), seed mass (SM), days to female flowering (DFF), and anthesis-silking interval (ASI).

The PAB population, derived from the cross between PA and PB, exhibited an intermediate pattern: FA1 (53.67%) correlated with GY, DFF, and SM, with GY and SM showing a negative relationship with DFF. FA2 (28.92%) was associated with GY and ASI, with high-performance progenies concentrated in Group 2. Genetically, FA1 (62.42%) maintained the association with these three traits, while FA2 (20.08%) stood out for

its linkage with ASI and SM, reinforcing the potential of FA1 for selecting productive progenies (Groups 1 and 2).

Conversely, the PBA population displayed a pattern similar to PA but with genetic influence from PB. Phenotypically, FA1 (73.45%) was linked to flowering (DFF), while FA2 (24.58%) was associated with yield (GY). In the genotypic analysis, FA1 correlated with DFF, ASI, and SM, whereas FA2 showed a strong correlation with GY, identifying Group 2 as the target for selecting high-yielding progenies.

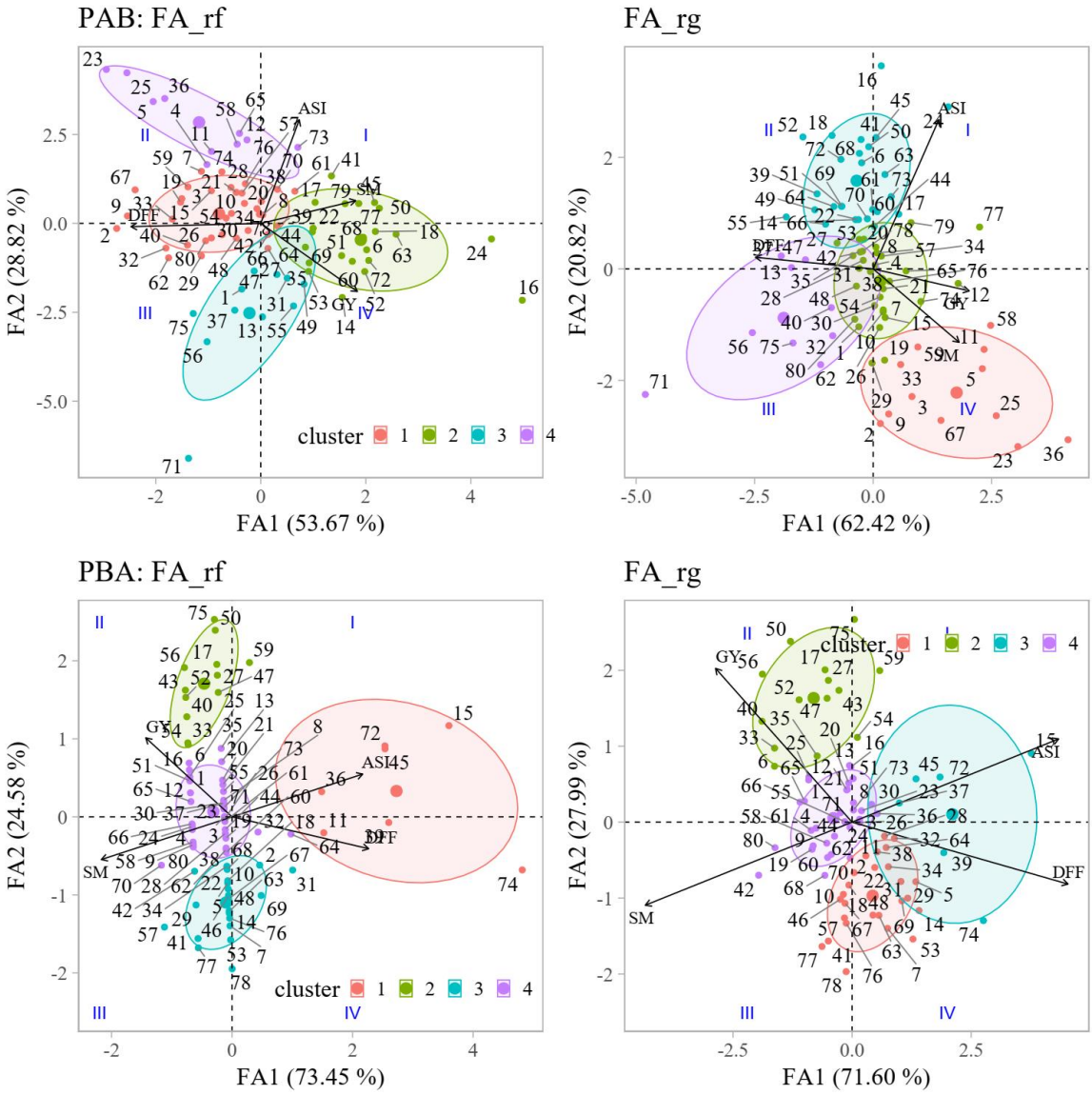


Figure 3 Factor analysis Based on phenotypic (rf) and genotypic (rg) Correlations in maize progenies from PAB and PBA intercross populations. The graphs display the projection of the progenies in the two-dimensional plane formed by the two factors (Factor), with distinct clusters (groups) identified by colors. Each progeny is represented

by numbers, and the ellipses outline the clusters based on grain yield (GY), seed mass (SM), days to female flowering (DFF), and anthesis-silking interval (ASI).

The results demonstrate a clear distinction between Factor FA1 (primarily associated with flowering traits) and Factor FA2 (associated with yield traits), enabling targeted selection strategies. Population PA exhibited an inverse relationship between yield and flowering traits. In contrast, PB showed a stronger influence of flowering in FA1 but retained productive potential in FA2.

The hybrid populations (PAB and PBA) reflected the combination of parental traits: PAB allowed the identification of genotypes with a balanced trade-off between yield and efficient flowering, whereas PBA maintained the PA pattern but with greater genetic variability.

These findings confirm that the parental populations (PA and PB) possess complementary traits transmitted to the hybrid populations (PAB and PBA), aligning with previous reports on heterosis and genetic complementarity (Baldauf et al. 2018). Earlier studies by Baldauf et al. (2018) emphasize that crosses between genetically distinct populations produce hybrids with increased variability and adaptive potential across diverse environments.

Populations PA and PB reflect classic recurrent selection patterns, as discussed by Almada et al. (2024), where progenies from short flowering cycles are often associated with high yield. Progenies with shorter anthesis-silking interval (ASI) and earlier days to flowering (DFF) tend to synchronize reproductive events more effectively, a critical factor in environments prone to water and heat stress (Li et al. 2023; Zhuang et al. 2024) and in interpopulation crosses, where male-female flowering synchrony can be decisive.

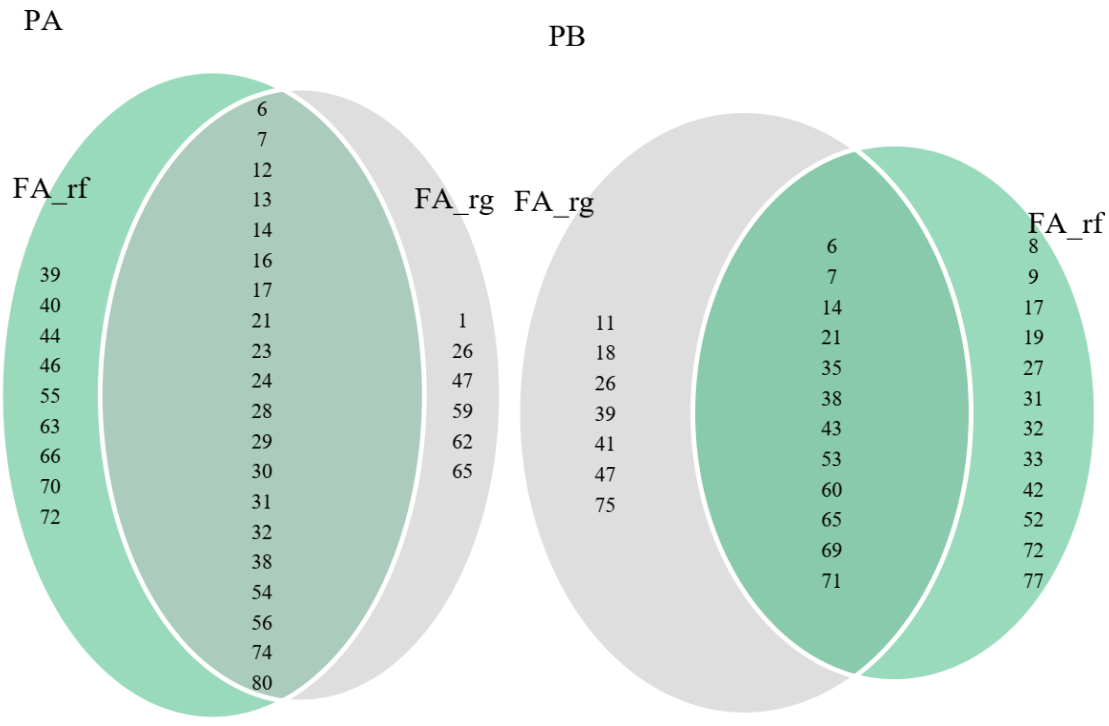
Phenotypic correlation-based analysis directly reflects observable traits and plant behavior in the environment. This is crucial for identifying genotypes that perform well under real field conditions, as noted by Cabral et al. (2011). They argue that traits that are genetically correlated but do not exhibit significant phenotypic correlation may not respond to selection since selection is based on phenotype.

On the other hand, according to Cruz et al. (2014), genetic correlation-based analysis provides insights into the heritable potential of genotypes, allowing the identification of traits that can be consistently passed on to future generations. Integrating these analyses enables more robust selection, considering both phenotypic and the genetic performance of

progeny. This facilitates informed decision-making, accelerating genetic gain. The separation into distinct groups allows breeders to conduct more targeted and efficient selection, focusing on progeny groups with desirable traits for the breeding program.

In Figure 4, the populations PA, PB, PAB, and PBA exhibited shared progenies selected through factor analysis based on phenotypic and genetic correlation, accounting for 57.10%, 38.70%, 65.50%, and 20.80% of the progenies, respectively. These progenies display simultaneous phenotypic and genetic expression for agronomic traits, positioning them as ideal candidates for breeding programs.

The exclusive progenies in each group, selected based on phenotypic performance, reflect a stronger influence of environmental and genetic effects, whereas those selected via genotypic correlation indicate the influence of additive genetic variation. Falconer (1996) emphasized that genotypic selection minimizes the impact of adverse environmental conditions, enhancing the heritability of selected genotypes. However, phenotypic selection remains relevant, particularly for adaptation to specific growing conditions (Cabral et al., 2011). Nevertheless, selecting progenies for both phenotypic and genetic expression increases the reliability of reproducibility in subsequent generations of maize breeding.



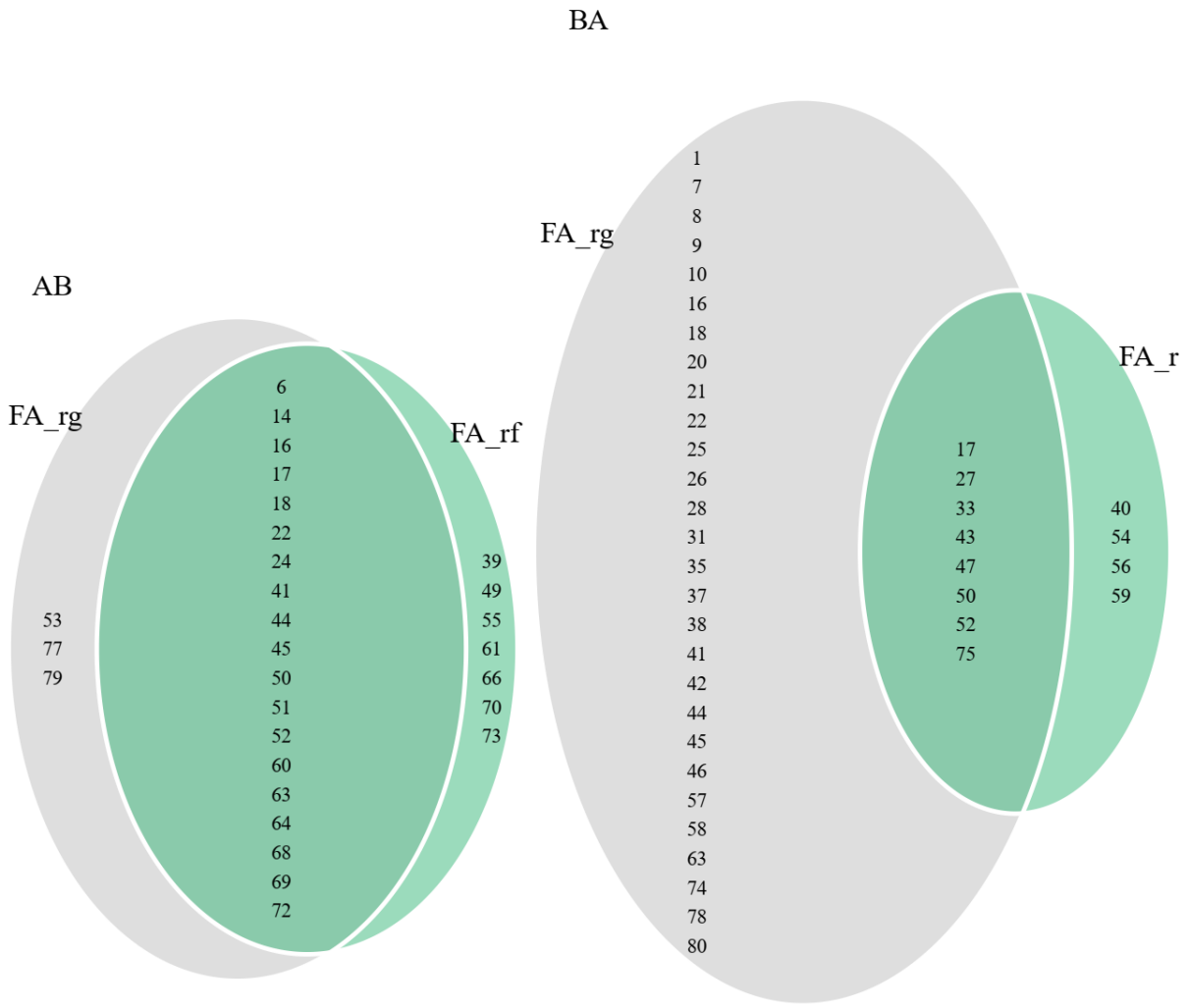


Figure 4 Common progenies simultaneously selected based on the phenotypic (rf) and genotypic (rg) expression of populations PA, PB, AB, and BA.

Conclusion

The PB population exhibits greater additive genetic variance potential for improving grain yield, seed mass, days to flowering, and anthesis-silking interval through the Bulk Pollen Pollination crossing method. Factorial analyses allowed the identification of corn progeny variability patterns based on phenotypic and genotypic correlations, optimizing the selection of high-performance progeny with simultaneous phenotypic and genetic expression in the reciprocal recurrent selection program.

References

Almeida PHS, Vilela VJB, Torres IY, Uberti A, Delima RO, Dos Reis EF. 2024. Genetic potential of maize populations derived from commercial hybrids for interpopulation breeding. Rev Caatinga. 37:1–7. https://doi.org/10.1590/1983-21252024v3711736rc

Baldauf JA, Marcon C, Lithio A, Vedder L, Altrogge L, Piepho HP, Schoof H, Nettleton D, Hochholdinger F. 2018. Single-Parent Expression Is a General Mechanism Driving Extensive Complementation of Non-syntenic Genes in Maize Hybrids. Curr Biol. 28(3):431-437.e4. https://doi.org/10.1016/j.cub.2017.12.027

Bartlett MS. 1951. The Effect of Standardization on a χ 2 Approximation in Factor Analysis. Biometrika. 38(3/4):337. https://doi.org/10.2307/2332580

BORÉM, A; GALVÃO, J. C. C.; PIMENTEL MA. 2017. No TitleMilho: do plantio à colheita. 2nd ed. Viçosa.

Cabral a. JEA. 2011. Análise de trilha do rendimento de grãos de feijoeiro (Phaseolus vulgaris L .) e seus componentes 1 Path analysis of grain yield of common bean (Phaseolus vulgaris L .) and its components. Rev Ciência Agronômica. 42(1):132–138.

Cruz, Cosme Damião; Carneiro, Pedro Crescencio souza; Regazzi AJ. 2014. Modelos biometricos aplicados ao melhoramento genetico. 3rd ed. UFV, editor. Vicosa: 2014.

Falconer DS. 1996. Introduction to quantitative genetics. Pearson Educ Índia,.

Hartigan, J. A., & Wong MA. 1979. Algorithm AS 136 A K-Means Clustering Algorithm. Appl Stat. 28(1):126-130.

Humberto Gonçalves dos Santos, Paulo Klinger Tito Jacomine, Lúcia Helena Cunha dos Anjos, Virlei Álvaro de Oliveira, José Francisco Lumbreras, Maurício Rizzato Coelho, Jaime Antonio de Almeida, José Coelho de Araújo Filho, João Bertoldo de Oliveira TJFC. 2018. No TitleSistema Brasileiro de Classificação de Solos. 5th ed. Embrapa Brasília, DF: Empresa Brasileira de Pesquisa Agropecuária Embrapa Solos.

Kassambara A, Mundt F. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Package Version 1.0.7. R Packag version.

Li H, Liu K, Li Z, Zhang M, Zhang Y, Li S, Wang X, Zhou J, Zhao Y, Liu T, Li C. 2023. Mixing trait-based corn (Zea mays L.) cultivars increases yield through pollination synchronization and increased cross-fertilization. Crop J. 11(1):291–300. https://doi.org/10.1016/j.cj.2022.05.007

Markus A, Grieder S, Revelle W, Auerswald M, Moshagen M, Ruscio J, Roche B, Lorenzo-seva U, Navarro-gonzalez D, Steiner MM. 2024. Package ‘ EFAtools .’

Olivoto T, Nardino M. 2021. Genetics and population analysis MGIDI : toward an effective multivariate selection in biological experiments. 37(December 2020):1383–1389. https://doi.org/10.1093/bioinformatics/btaa981

Pedro C, Marçola MA, Charimba AM, de Queiroz LGC, de Souza JC. 2023. Genetic potential of maize full-sib progenies subjected to a reciprocal recurrent selection. Pesqui Agropecu Bras. 58:1–11. https://doi.org/10.1590/S1678-3921.pab2023.v58.03134

Ramalho, M. A. P., Abreu, A. F. B., Santos, J. B., & Nunes JA. 2012. Aplicações da genética quantitativa no melhoramento de plantas autógamas. 1st ed. Lavras: UFLA.

Rocha JR do AS de C, Machado JC, Carneiro PCS. 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

Silva-Díaz R, García-Mendoza P, Faleiro-Silva D, De Souza CL. 2018. Determinación de componentes de la varianza y parámetros genéticos en una población segregante de maíz tropical. Bioagro. 30(1):67–77.

Talabi AO, Badu-Apraku B, Fakorede MAB. 2017. Genetic variances and relationship among traits of an early maturing maize population under drought-stress and low nitrogen environments. Crop Sci. 57(2):681–692. https://doi.org/10.2135/cropsci2016.03.0177

Wang P, Zhang H, lyle D, Li D, Wang G, Pan Q, Wang J. 2019. Bulk pollen pollination in

maize for efficient construction of introgression populations with high genome coverage. Plant Breed. 138(3):252–258. https://doi.org/10.1111/pbr.12684

Zhuang L, Wang C, Hao H, Song W, Guo X. 2024. Maize Anthesis-Silking Interval Estimation via Image Detection under Field Rail-Based Phenotyping Platform. Agronomy. 14(8). https://doi.org/10.3390/agronomy14081723

ARTIGO 3- Multilevel yield prediction of half-sib corn progenies derived from intra- and interpopulation crosses

Periódico: Euphytica, versão preliminar.

César Pedro [a] (https://orcid.org/0000-0002-2963-8652) . João Cândido de Souza [a*] (https://orcid.org/0000-0001-9580-4631)

[a] *Universidade Federal de Lavras, Departamento de Biologia, Aquenta Sol, Lavras - MG, CEP 37200-900, Brasil.*

*e-mail: cansouza@ufla.br (autor correspondente)

Abstract This study evaluated the contribution of crosses and populations to the yield and genetic variability of corn half-sib progenies using multilevel models (MLM). Four experiments were conducted between March and July 2024 at the Scientific and Technological Development Center for Agriculture at the Federal University of Lavras, located in southern Minas Gerais State, Brazil. 312 half-sib progenies were evaluated (78 from each population), derived from intrapopulation crosses (A and B) and interpopulation crosses (AB and BA). Two commercial double-cross hybrids were used as checks in an alpha-lattice design (10 × 8) with two replications. Progeny prediction was performed using MLM, and validated through comparison with Ordinary Least Squares Regression (OLS) and Random Forest (RF). The MLM more effectively captured the genetic variability of the progenies within crosses and populations. Interpopulation crosses accounted for higher heterosis and progeny yield. Compared to OLS and RF, MLM demonstrated high-performance in explaining the genetic variability of progenies across hierarchical structure levels. Multilevel models proved to be a highly predictive tool, supporting the accurate selection of high-yielding progenies. These findings reinforce the relevance of MLM in optimizing strategies for corn genetic advancement.

Keywords *Zea mays* L. Multilevel models . Ordinary Least Square Regression . Random Forest

Predição multinível de progenies meio irmãos de milho derivados de cruzamentos intra- e interpopulacionais.

Resumo: Objectivou-se avaliar a contribuicao dos cruzamentos e populacoes na produtividade e variabilidade genetica de progênies meio-irmãos de milho, utilizando modelos multiníveis (MLM). Quatro experimentos foram conduzidos entre março e julho de 2024, no Centro de Desenvolvimento Científico e Tecnológico da Agricultura da Universidade Federal de Lavras, localizada no sul do Estado de Minas Gerais, Brasil. Foram avaliadas 312 progênies meio-irmãs (78 de cada população derivadas de cruzamentos intrapopulacionais: A e B, e de cruzamentos interpopulacionais: AB e BA. Foram utilizados dois híbridos duplos comerciais como testemunhas, em delineamento alpha-lattice (10 × 8), com duas repetições. A predicao das progenies foi feita por meio de modelos MLM validadas por comparacao com Ordinary Least Squares Regression (OLS) e Random forest (RF). Os MLM capturaram de forma mais eficiente a variabilidade genética das progênies dentro de cruzamentos e populacoes. Os cruzamentos interpopulacionais explicam a maior heterose e produtividade das progenies. Os

MLM superaram OLS e RF, na contribuicao da variabilidade genética das progenies em diferentes niveis de estrutura hierárquica. Os modelos multiníveis destacaram-se como ferramentas com maior poder preditivo e suporte à seleção acurada de progênies superiores. Esses resultados reforçam a relevância dos modelos multiníveis na otimização de estratégias para o avanço genético do milho.

Palvras-chave: *Zea mays* L. . Modelos multinível . Ordinary Least Square Regression . Random Forest

Introduction

Corn (*Zea mays* L.) breeding faces the critical challenge of maximizing agricultural yield to meet growing global demands for food security and sustainability. In this context, Reciprocal Recurrent Selection (RRS) stands out as one of the most effective strategies, leveraging genetic variability through intra- and inter-population crosses that promote continuous gains in progeny performance (Dos Reis et al. 2014; Yong et al. 2019; Pedro et al. 2023). However, the success of this method fundamentally depends on the ability to accurately predict variability and the yield potential of progeny, considering the hierarchical structure of breeding programs, which encompasses levels of crosses, populations and progeny.

To address this, the development of robust statistical models capable of effectively capturing these relationships while minimizing bias and improving selection accuracy is essential. Multilevel Models (MLM) emerge as a promising solution due to their intrinsic ability to analyze hierarchically structured data, overcoming limitations of traditional techniques such as Ordinary Least Squares (OLS) regression, which fails to adequately account for dependencies among grouped observations (Hoffman and Walters 2022).

MLMs have been widely applied in several areas of knowledge, with emphasis on health, education and psychology, economics and social sciences (Asampana Asosega et al. 2024). In agriculture, although less explored, recent studies have demonstrated their effectiveness in agronomic predictions for different contexts, such as Hoang et al. (2020) in spatial modeling of rice yield and Li et al. (2020) in wheat quality assessment. However, in the specific context of corn breeding, particularly in RRS programs, the application of these models remains limited, representing a significant scientific gap.

This study addresses this gap through two primary objectives: assess the contribution of different crossing strategies (intracross and intercross) and populations to grain yield variability in corn progeny using Multilevel Models (MLM). Compare the predictive performance of MLM against traditional methods (Ordinary Least Squares: OLS and Random Forest: RF) for predicting grain yield of corn progeny.

The expected outcomes include identifying key hierarchical factors influencing yield and validating the superiority of MLMs over OLS and RF approaches. These findings will have significant practical implications, providing insights for developing more efficient RRS strategies that enhance corn breeding with greater precision and sustainability.

Material and methods

First assessment of corn populations

From November 2021 to March 2022, 102 full-sib corn progenies from the eighth cycle of reciprocal recurrent selection (RRS) were evaluated. These progenies were derived from two populations: A (DKB 333B) and B (DOW 657), as part of the corn breeding program at the Federal University of Lavras (UFLA). The trials were conducted at the experimental area of Muquém Farm (21°12'S, 45°59'W, altitude 918.84 m), located at the Center for Scientific and Technological Development in Agriculture (CDCTA/UFLA), in Lavras, Minas Gerais (MG), Brazil. The experimental design was an alpha-lattice with three replications and 24 blocks. Among these 102 progenies, 15% were selected based on genetic divergence and agronomic performance (Pedro et al. 2023). Subsequently, in November 2022, the selected progenies were planted into heterotic groups A and B, followed by controlled intercrossing within each group in a full diallel mating scheme. The resulting progenies from intercrossing within each heterotic group were bulk-harvested, forming two distinct populations (A and B).

From November 2023 to January 2024, a total of 2,000 plants (1,000 from each population, A and B) were cultivated in two 10 × 10 m blocks, spaced two meters apart. From these populations, four intra- and interpopulation half-sib progeny populations were developed using the Bulk Pollen Pollination (BPP) method (Wang et al. 2019). To generate intrapopulation progenies (A and B), pollen from 10 randomly selected plants was collected, mixed, and used to pollinate the lower ear of a plant from the same population. The same procedure was applied to population B. For interpopulation progenies, pollen from population A was used to pollinate plants from population B, and vice versa. As a result, four half-sib progeny populations were established: two from intracross (A and B) and two from intercross (AB and BA) populations. From each population, 78 progenies exhibiting good morphologic traits, well-filled ears, and sufficient kernel set on both ears were selected, totaling 312 progenies for experimental evaluation (Fig. 1).

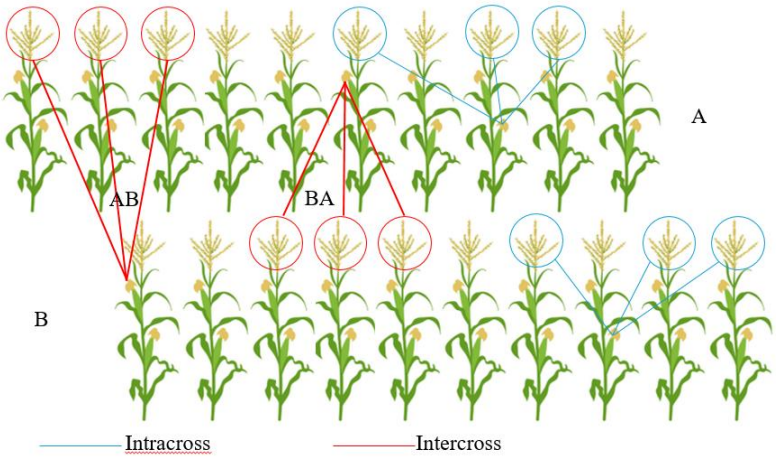


Fig. 1 Breeding scheme for intracross populations (A and B) and intercross populations (AB and BA) using the Bulk Pollen Pollination method.

Second assessment of corn populations

The 312 half-sib progenies and two commercial double-cross hybrids (checks) were evaluated in an alpha-lattice design (10 × 8) with two replications in four experiments (78 each). Sowing and harvesting took place between March 5 to 6, 2024, and June 2024, at a density of four seeds per linear meter in 4-meter-long plots spaced 0.6 m apart. The trials were conducted at the same location as the first evaluation, where the experimental area soil is classified as Red-Yellow Latosol (Oxisol) with gently undulating topography, belonging to the Ferralsols group according to the international taxonomic classification (FAO 2014). The climate is classified as rainy temperate (Cwa) under the Köppen-Geiger system, with an average temperature of 22.96 °C and cumulative rainfall of 218.4 mm during the experimental period, as recorded by the Lavras weather station (21°13'34.0"S, 44°58'47.0"W).

Basal fertilization at planting consisted of 250 kg ha^{-1} of fertilizer (8% N, 28% P_2O_5, and 16% K_2O). A topdressing application of 200 kg ha^{-1} of urea-N (45% N) was performed 25 days after sowing. Crop management followed region-specific recommendations (Borém, 2017).

Evaluated traits and data analysis

The evaluated traits included grain yield (GY, t ha^{-1}), seed mass (SM, g), and days to female flowering (DFF). Data analysis was based on ordinary least squares (OLS) models, Random Forest (RF), and Multilevel Models (MLM). The latter considered three hierarchical levels of random variance (Crosses, Populations, and Progenies).

MLM1: Progeny nested within crosses

$$GY_{ij} = \beta_0 + \beta_1 \text{ x } SM_{ij} + \beta_2 \text{ x } DFF_{ij} + u_j + u_{ij} + \epsilon_{ij}$$

Where: GY_{ij}: Grain yield for the *i*-th progeny in the *j*-th cross; β_0: Overall intercept; β_1, β_2: Regression coefficients for SM e DFF, respectively; u_j: Random effect of the j -th cross, $u_j \sim N(0, \sigma^2_{cross})u_{ij}$: Random effect of the *i*-th progeny within the j -th cross, $u_{ij} \sim N\left(0, \sigma^2_{Progenie(Cross)}\right)$. ϵ_{ij}: Residual error, $\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$

MLM2: Progeny nested within populations

$$GY_{ik} = \beta_0 + \beta_1 \text{ x } SM_{ik} + \beta_2 \text{ x } DFF_{ik} + u_k + u_{ik} + \epsilon_{ik}$$

GY_{ik}: Grain yield for the i- th progeny in the k -th population. u_k: Random effect of the k-th population, $u_k \sim N\left(0, \sigma^2_{Population}\right)$. u_{ik}: Random effect of the i- th progeny within the k -th population, u_{ik}~ $\text{N}(0, \sigma^2_{Progenie(Population)})$.

MLM3: Progeny nested within populations and crosses

$$GY_{ijk} = \beta_0 + \beta_1 \text{ x } M_{ijk} + \beta_2 \text{ x } DFF_{ijk} + u_j + u_k + u_{ijk} + \epsilon_{ijk}$$

Where**:** GY_{ijk}: Grain yield for the i -th progeny in the j -th population and k -th cross. u_j: Random effect of the j, $u_j \sim N(0, \sigma^2_{cross})$. u_k: Random effect of the k -th population, $u_k \sim N(0,$ σ²_Population). u_{ijk}: Random effect of the i -th progeny within the cross-population combination, $u_{ijk} \sim N\left(0, \sigma^2_{Progenie(Popuation/cross)}\right)$.

The variance components estimated using the nlme package were:

σ^2_ϵ = Residual variance (error);
$\sigma^2_{Progenie/\text{Populations}}$ = Variance associated with the Progenies within Populations level;
$\sigma^2_{Progenie/\text{Cross}}$ = Variance associated with the Progenies within Cross level;
$\sigma^2_{\text{Population/Cross}}$ = Variance associated with the Populations within Cross level;
$\sigma^2_{Progenie/\text{Population/Cross}}$ = Variance associated with the Progenies within the Populations and Crosses level;
$\sigma^2_{\text{Population}}$ = Variance associated with the Populations level;
σ^2_{Cross} = Variance associated with the Cross level.

Multilevel Model (MLM) Metrics

The marginal and conditional R² metrics were used to evaluate the proportion of variance explained by the fixed effects and fixed + random effects of the model, respectively, as proposed by Johnson (2014).

$$R^2_{Marginal} = \frac{\sigma^2_{fixed}}{\sigma^2_{fixed} + \sigma^2_{random} + \sigma^2_\epsilon}$$

$$R^2_{Conditional} = \frac{\sigma^2_{fixed} + \sigma^2_{random}}{\sigma^2_{fixed} + \sigma^2_{random} + \sigma^2_\epsilon}$$

Where: σ^2_{fixed}: Variance explained by fixed variables (fixed effects). σ^2_{random}: Combined variance explained by random components. σ^2_ϵ: Residual variance (error)

Intraclass correlation coefficient (ICC)

The intraclass correlation coefficient (ICC) was used to measure the proportion of total variance attributed to the random effect components of the multilevel models (MLM).

- MLM1: $ICC_{crosses} = \frac{\sigma^2_{Cross}}{\sigma^2_{cross} + \sigma^2_{Progenie(cross)} + \sigma^2_\epsilon}$
- MLM2: $ICC_{Populations} = \frac{\sigma^2_{Population}}{\sigma^2_{Population} + \sigma^2_{Progenie(Population)} + \sigma^2_\epsilon}$
- MLM3:

$$ICC_{crosses} = \frac{\sigma^2_{Cross}}{\sigma^2_{cross} + \sigma^2_{Population} + \sigma^2_{Progenie(Population/cross)} + \sigma^2_{\epsilon}}$$

$$ICC_{Populations} = \frac{\sigma^2_{Population}}{\sigma^2_{cross} + \sigma^2_{Population} + \sigma^2_{Progenie(Population/cross)} + \sigma^2_{\epsilon}}$$

The calculations of R^2, ICC, and variance decomposition were performed using the following packages: *lme4* for fitting the multilevel models, *MuMIn* for computing marginal and conditional R^2, and *performance* for extracting the ICC.

The contribution of variance from random levels to the model was estimated as:

$$\text{Contribuition } (\%) = \left(\frac{\sigma^2_{\text{level}}}{\sigma^2_{\text{total}}}\right) \times 100$$

Ordinary Least Square (OLS)

Ordinary Least Squares (OLS) was used for multiple regression analysis, estimating the coefficients of the relationship between the dependent and independent variables according to the equation:

$$Y = \beta_0 + \beta_1 \text{ x } SM + \beta_2 \text{ x } DFF + \epsilon$$

Where: Y is grain yield, SM and DFF are the explanatory variables, $\beta_0, \beta_1, \beta_2$ are the coefficients to be estimated, ϵ is the random error.
The model was fitted using the stats package.

Random Forest (RF)

The Random Forest (RF) model is a machine learning algorithm used to predict grain yield based on independent variables. The data were initially split into a training set (70%) to fit the RF model, where 300 decision trees were constructed. Final predictions were obtained by averaging the predictions of all trees, which were built using bootstrap samples and random subsets of explanatory variables. The model fitting process involved random variable selection at each node split and the construction of deep trees to capture nonlinear interactions among variables.

A validation set (15%) was used to fine-tune the model during the fitting process, aiding in the selection of hyperparameters such as the number of trees and maximum depth. Meanwhile, a test set (15%) was reserved for the final evaluation of the model's performance on unseen data.

Dummy coding of variables and Stepwise regression based in OLS

The variables 'Crossing' and 'Population' were converted into binary variables for use in the OLS and RF models, with independent variables selected via forward stepwise regression.

Model performance evaluation

The performance of the models was evaluated based on the following metrics:

Coefficient of determination (R^2): Measured as the proportion of variance in observed values explained by the model's predicted values:

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

Where: y_i: : Observed value, $\hat{y}_i$: Predicted value from the model, $\bar{y}$: Mean of observed values, n: Total number of observations.

Root Mean Square Error (RMSE): Used to assess the magnitude of error between observed and predicted values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$$

Where: n: Number of observations, Y_i: Observed values, $\hat{Y}_i$: Predicted values.

Progeny averages and heterosis

Heterosis was estimated by comparing the AB and BA hybrids with their respective parental populations, A and B. First, the average yields of the parental populations (A and B) and the interpopulation hybrids (AB and BA) were obtained. Heterosis was calculated as the difference between the average yield of the hybrids (AB or BA) and the average yield of the corresponding parental populations (A and B). The absolute value of this difference was then divided by the mean of the parental values and expressed as a percentage, according to the equation:

$$\text{Heterosis (\%)} = \left(\frac{\text{GY}_{\text{Hybrid}} - \text{GY}_{\text{Mean of parents}}}{\text{GY}_{\text{mean od parents}}}\right) \times 100$$

$\text{GY}_{\text{Hybrid}}$ is the hybrid yield (AB or BA). $\text{GY}_{\text{mean of parents}}$ is the average yield of the parental populations A and B. All analyses were performed using the R software (R Core Team 2023).

Results and Discussion

The results of the multilevel modeling (MLM) highlighted the impact of genetic variation sources across three hierarchical random levels: crosses, populations and progenies. The fixed effects (SM and DFF) were also significant for predicting grain yield, with SM exerting a positive effect and DFF a negative one, indicating that increased SM and reduced DFF tend to predict the yield performance of corn populations positively.

In the MLM1 model, most of the variation was attributed to differences among progenies within crosses, while crosses themselves had a smaller contribution, although both showed significant effects. In MLM2, most of the variation was explained by differences among progenies within populations, whereas populations had a lesser influence, though still with significant effects. In MLM3, most of the yield variance was explained by progenies within populations and crosses. However, this last model masked the effect of crosses, highlighting the consistent and significant contribution of populations to yield variability, as also observed in MLM2. Nevertheless, the findings underscore the importance of crosses, which introduce additional heterogeneity confirmed in MLM1 and are crucial in explaining part of the genetic variability in progeny yield in corn RRS.

The intraclass correlation coefficient (ICC) showed little variation among models and highlighted the high proportion of variability explained by progenies within crosses, populations, or both. The conditional R^2 indicated that fixed and random effects jointly accounted for most of the variability, though the magnitude exhibited only minor variation across models. The residual variance in MLM2 and MLM3 was higher than in MLM1, likely due to the increased complexity introduced by populations to explain corn yield variability.

The results confirm the influence of fixed predictors on yield, aligning with studies by various authors (Pedro et al. 2023; Dube et al. 2023; Almeida et al. 2024), which emphasize the effect of SM and DFF traits on GY improvement. MLM1 demonstrated the extent to which crosses contribute to significant genetic variability in corn progenies, corroborating Mukri et al. (2022) who highlighted the importance of crosses in maximizing heterosis. Studies such as those by Dube et al. (2023) and Li et al. (2021) also reveal substantial genetic variation among progenies within populations.

In MLM3, the simultaneous inclusion of crosses and populations showed that populations have a greater overall impact, though crosses broaden the available genetic variability for selection. These findings suggest that breeding programs should prioritize selecting progenies within populations and crosses to fully exploit existing variability while avoiding genetic base narrowing. The use of hierarchical models, as in this study, enables more precise experimental design, aiding in identifying critical sources of genetic variation for enhancing corn yield in SRR.

Table 1 Effects of SM, DFF, Cross, Populations, and progenies on corn grain yield

Predictors	**MLM1**	
	Estimates	**CI**
(Intercept)	14	9.32 – 18.68 ***
SM	0.05	0.01 – 0.09 **
DFF	0.15	-0.21 – -0.08 ***
Random effects	**_**	**LRT (χ2)**
σ^2	2.37 (42.63 %)	_
$\sigma^2_{Progenie/Cross}$	2.66 (47.84 %)	102.39 ***
σ^2_{Cross}	0.53 (9.53 %)	14.38 ***
ICC	0.57	_
Marginal R^2 / Conditional R^2	0.055 / 0.598	_
Predictors	**MLM2**	
(Intercept)	15.36	10.52 – 20.19 ***
SM	0.07	0.03 – 0.11 **
DFF	-0.18	-0.25 – -0.11 ***
Random effects	**_**	**LRT (χ2)**
σ^2	2.38 (44.32 %)	_
$\sigma^2_{Progenie/Populations}$	2.48 (46.17 %)	92.86 ***
$\sigma^2_{Populations}$	0.51 (9.50 %)	23.64 ***
ICC	0.56	_
Marginal R^2 / Conditional R^2	0.089 / 0.597	_
Predictors	**MLM3**	
(Intercept)	15.27	10.42 – 20.13 ***
SM	0.07	0.03 – 0.11 ***
DFF	-0.18	-0.25 – -0.11 ***
Random effects	**_**	**LRT (χ2)**
σ^2	2.38 (43.73 %)	_
$\sigma^2_{Progenie(Populations/Cross)}$	2.48 (45.59 %)	92.98 ***
$\sigma^2_{Populations/Cross}$	0.38 (7.02 %)	9.41 **
σ^2_{Cross}	0.20 (3.67 %)	0.14
ICC	0.56	_
Marginal R^2 / Conditional R^2	0.086 / 0.600	_
$N_{Progenies}$ / $N_{Populations}$/ N_{Cross}	312/4/2	_
Observations	624	_

ICC = intraclass correlation coefficient

Table 2 Ordinary Least squares (OLS) estimates using Backward Stepwise Dummy predictors (DBS) method and predictor importance from Random Forest (RF).

Predictors	DBS+OLS	RF
	Estimates	Importance
(Intercept)	17.17 ***	
SM	0.08 ***	1.64
DFF	-0.20 ***	1.48
Cross Intracross	-1.77 ***	1.37
Pop AB	-0.71 **	1.04
Pop B	1.22 ***	1.05

Among the MLM models, MLM3 was selected to compare the predictive ability of OLS and RF models in assessing the variability of corn progeny yield, using R^2 and RMSE metrics (Fig. 2). The choice of MLM3 is justified by the need to capture the yield potential of progenies within each population and cross.

The results confirm that SM and DFF are significant predictors of yield in both models. Among the dummy variables, the negative coefficient of the predictor Cross_Intracross indicates that progenies from intrapopulation crosses exhibit a significant reduction in yield compared to those from interpopulation crosses. Meanwhile, the variable Pop AB shows a negative coefficient, suggesting low performance relative to population BA, whereas Pop B displays a positive coefficient, indicating higher yield compared to Pop A.

Predictor importance analysis revealed Cross_Intracross and SM as the most relevant variables for OLS and RF. This consistency underscores the strong relationship between these variables and yield.

Additionally, the results confirm the superiority of the multilevel model (MLM) in predicting yield, outperforming both OLS and RF. The enhanced performance of MLM, evidenced by higher R^2 values and lower RMSE, stems from its ability to model the hierarchical structure of the data, enabling nested exploration of the genetic variability of corn yield in SRR.

Comparing with OLS, RF demonstrated better performance across all populations, though it remained low to MLM. The limited performance of OLS aligns with findings from Li et al. (2019) and Zhu et al. (2021) who highlighted the inefficiency of OLS models in capturing the complexity of hierarchically structured data. Although OLS models often perform well in most crops, they tend to underestimate variability, leading to less accurate predictions (Liu et al. 2021).

When comparing this study to those of Li et al. (2019) and Zhu et al. (2021) which analyzed wheat yield prediction using MLM and OLS, the consistent superiority of MLM in capturing structural variability is evident. However, when assessing the performance of RF, a widely used model as demonstrated by Baio et al. (2023) and Asamoah et al. (2024) the results are more contrasting. Baio et al. (2023) showed that RF achieved high accuracy in predicting corn yield using environmental variables, while Asamoah et al. (2024) emphasized its effectiveness in predicting nutrient use efficiency. Nevertheless, in the context of this study, RF proved ineffective in capturing the genetic variability of progeny yield within a multilevel genetic structure.

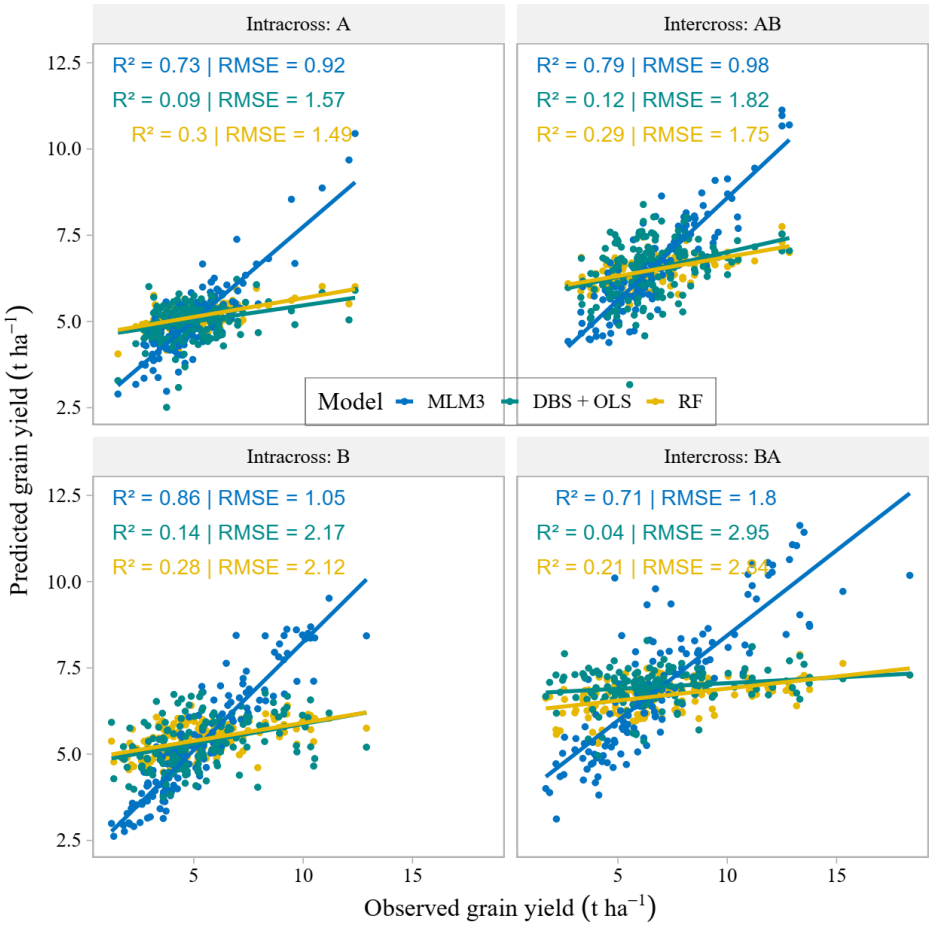


Fig. 2 Relationship between observed and predicted values from Multilevel model (MLM3), Random forest (RF) and Ordinary least squares based on Dummy Backward Stepwise (DBS+OLS). MLM3 = Multilevel models involving progenies within populations and crosses, R^2 = Coefficient of determination, RMSE = Root Mean Square Error.

The yield potential of progenies from different corn populations and crosses is shown in Fig. 3. The interpopulation crosses (intercross: AB and BA) exhibited the highest predicted yield averages (6.49 t ha^{-1} and 6.83 t ha^{-1}, respectively), attributed to heterotic effects resulting from the genetic complementarity between populations A and B. In contrast, intrapopulation crosses (intracross) in the parental populations A and B resulted in the lowest yield averages, with A being the least productive (5.01 t ha^{-1}), followed by B (5.25 t ha^{-1}). Nevertheless, population B stood out for having the highest genetic variability, indicating its relevance as a genetic base.

These findings align with the literature, as demonstrated by Wang et al. (2017) who highlighted the potential of interpopulation crosses to increase genetic variability and enable the selection of superior progenies. The heterotic effect observed in populations AB and BA confirms the role of interpopulation crosses in enhancing progeny productivity, as also reported by Reis et al. (2014), who emphasized heterosis as a key strategy in corn breeding. According to Cruz et al. (2014) the agronomic performance of genotypes, combined with genetic

divergence, is essential for maximizing gene complementarity and heterosis. Meanwhile, Reis et al. (2004) stressed the importance of diverse genetic recombination to broaden genetic potential.

Reis et al. (2014) observed in crosses between divergent base populations A and B a heterosis of 12.30% in cycle 0 and 24.90% in the third selection cycle. By the eighth cycle of these improved populations corresponding to this study heterosis gains ranged from 26.56% to 33.10% compared to the base population, demonstrating the potential of interpopulation crosses in enhancing variability and population means.

Finally, the results contrast with studies such as Bernardo (2001) who reported higher yield in intracross populations compared to intercross populations, which was not observed in populations A and B in this study. These discrepancies may be attributed to study duration, population type, crossing methodologies, climate, population origin, and other controlled or uncontrolled factors.

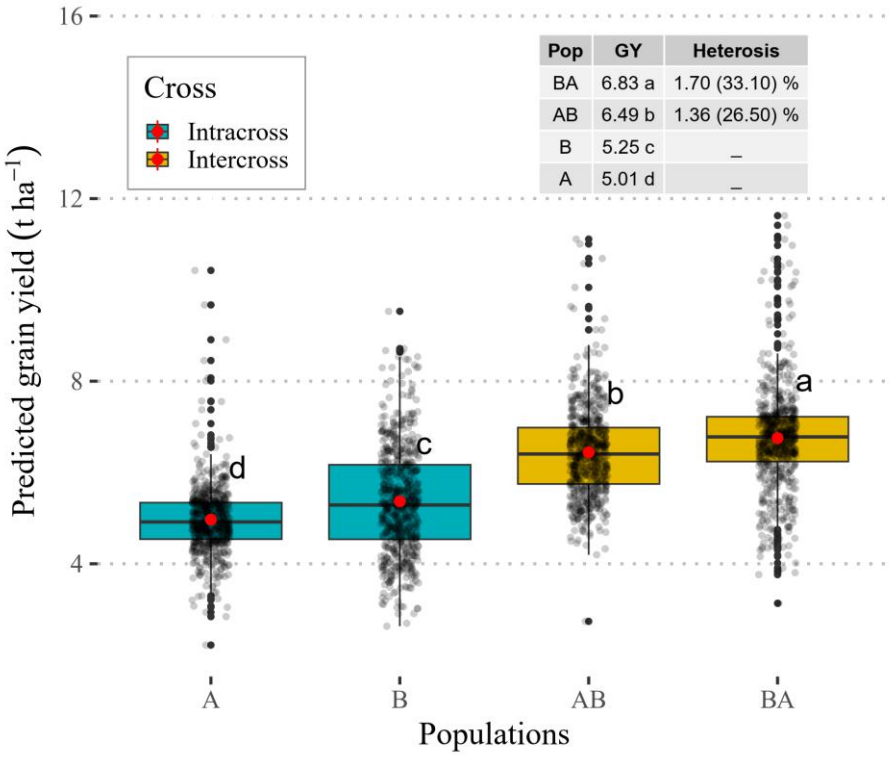


Fig. 3 Comparison of predicted grain yield (GY, kg ha^{-1}) and variability among intra- and intercrossed corn populations. Variability is represented by boxplots, with individual points indicating predicted GY values.

Based on the results from Fig. 4, it is possible to identify progenies with high-performance within and across breeding populations, considering the selection of the top 20 best-performing progenies. Significant differences were observed within each population, as determined by the Scott-Knott clustering test at a 5% probability level.

The progenies with the highest predicted yield should be prioritized for recombination. In intracross population A, the highest-yielding progenies exhibited values ranging from 6.25 t ha^{-1} to 10.04 t ha^{-1}, making them

ideal candidates for inclusion in the breeding program. Selecting the best progenies facilitates the accumulation of favorable alleles, promoting genetic gains in future generations.

In Intracross Population B, the predicted yield was higher than in Population A, ranging from 8.35 t ha^{-1} to 9.09 t ha^{-1}. Although these progenies demonstrate high-yielding, recombining the best progenies from different populations can enhance genetic variability, which is essential for the long-term sustainability of the breeding program

The interpopulation crosses yielded promising results. In the AB intercross population, yields ranged from 8.20 t ha^{-1} to 10.87 t ha^{-1}, while the BA intercross population achieved higher values, exceeding 9.00 t ha^{-1} and reaching up to 11.50 t ha^{-1}. These crosses highlight the heterotic effect, which is critical for developing high-yielding hybrids. The recombination of these progenies can significantly increase genetic variability and yield, advancing recurrent selection progress. According to Reis et al. (2014), these results indicate genetic divergence between parental populations and the presence of dominance, essential conditions for heterosis expression.

The selection intensity of progenies for recombination and evaluation has major implications in a breeding program. Interpopulation crosses (AB and BA), with higher genetic variability, allow for lower selection intensity, enhancing genetic diversity and the likelihood of developing superior hybrids. Kutka and Smith (2007) discuss the optimal number of genotypes to intercross, recommending at least five pure lines to form parental groups. Masoni et al. (2020) suggested crossing elite lines with local populations to generate variability, while Döring et al. (2015) emphasized that smaller sets of high-performing lines can optimize yield. Cruz et al. (2014) added that parents should be genetically divergent and potentially agronomic to maximize genetic gains.

The observed variability between populations underscores the importance of interpopulation crosses as a key strategy for developing new hybrids, promoting high heterosis levels, greater yield gains, and improving RS program efficiency.

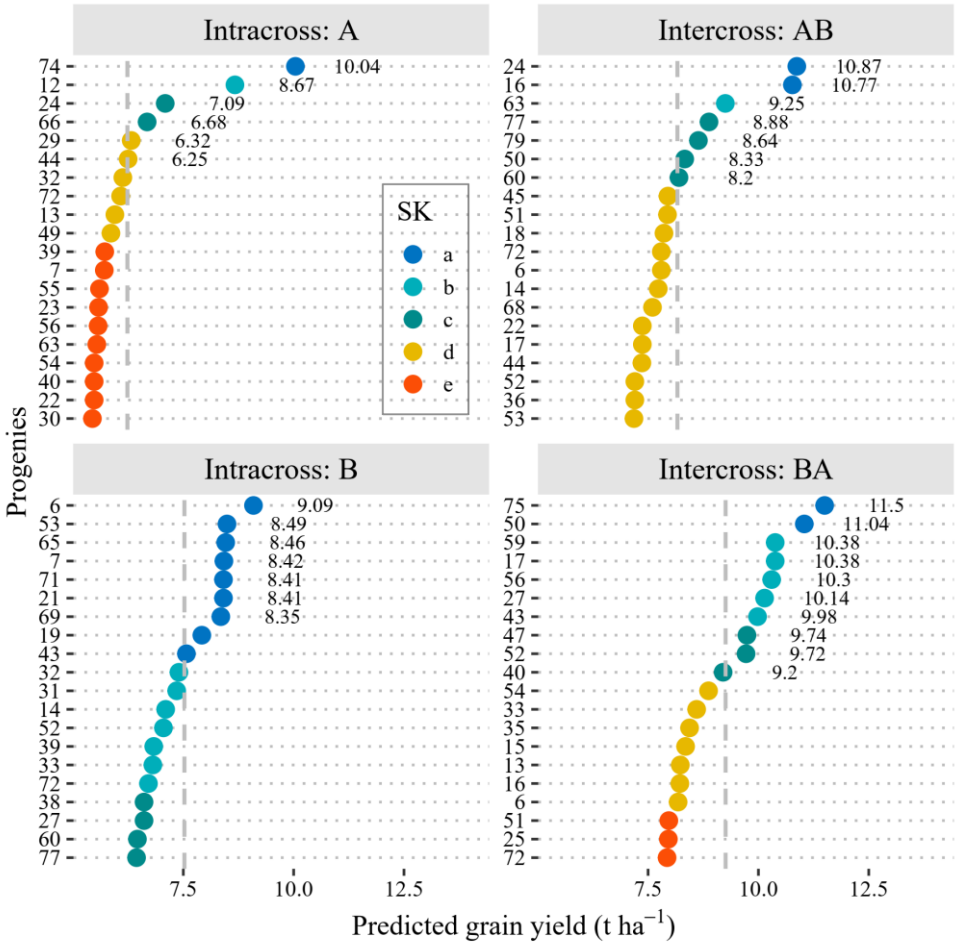


Fig. 4 Ranking of Intra- and Intercrossed Progenies from corn Populations Based on Predicted Grain Yield.

Conclusion

The genetic variability among progenies at the population level was the primary source of variability from the crosses, highlighting the importance of selecting high-performnce progenies within each population and cross. Interpopulation crosses are effective strategies for advancing corn breeding programs to maximize the genetic variability of progeny yield in reciprocal recurrent selection. The multilevel models outperformed Ordinary Least Squares and Random Forest in predicting grain yield. Multilevel models proved to be a better and essential tool for predicting progeny yield variability within the hierarchical context of populations and crosses in the corn reciprocal recurrent selection scheme.

Declarations

“This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, finance code 001), Brazil, for supporting this research.”

“The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.”

Conflict of interest The authors declare no competing interests.

Data availability

The data utilised in this study are available from the corresponding author upon a reasonable request.

References

Almeida PHS, Vilela VJB, Torres IY, et al (2024) Genetic potential of maize populations derived from commercial hybrids for interpopulation breeding. Rev Caatinga 37:1–7. https://doi.org/10.1590/1983-21252024v3711736rc

Asamoah E, Heuvelink GBM, Chairi I, et al (2024) Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana. Heliyon 10:e37065. https://doi.org/10.1016/j.heliyon.2024.e37065

Asampana Asosega K, Adebanji AO, Aidoo EN, Owusu-Dabo E (2024) Application of Hierarchical/Multilevel Models and Quality of Reporting (2010-2020): A Systematic Review. Sci World J 2024:. https://doi.org/10.1155/2024/4658333

Baio FHR, Santana DC, Teodoro LPR, et al (2023) Maize Yield Prediction with Machine Learning, Spectral Variables and Irrigation Management. Remote Sens 15:. https://doi.org/10.3390/rs15010079

Bernardo R (2001) Breeding potential of intra- and interheterotic group, crosses in maize. Crop Sci 41:68–71. https://doi.org/10.2135/cropsci2001.41168x

BORÉM, A; GALVÃO, J. C. C.; PIMENTEL MA (2017) No TitleMilho: do plantio à colheita., 2nd edn. Viçosa

Cruz, Cosme Damião; Carneiro, Pedro Crescencio souza; Regazzi AJ (2014) Modelos biometricos aplicados ao melhoramento genetico, 3rd edn. 2014, Vicosa

Döring TF, Annicchiarico P, Clarke S, et al (2015) Comparative analysis of performance and stability among composite cross populations, variety mixtures and pure lines of winter wheat in organic and conventional cropping systems. F Crop Res 183:235–245. https://doi.org/10.1016/j.fcr.2015.08.009

Dos Reis MC, Pádua JMV, Abreu GB, et al (2014) Estimates for genetic variance components in reciprocal recurrent selection in populations derived from maize single-cross hybrids. Sci World J 2014:. https://doi.org/10.1155/2014/540152

Dube SP, Sibiya J, Kutu F (2023) Genetic diversity and population structure of maize inbred lines using phenotypic traits and single nucleotide polymorphism (SNP) markers. Sci Rep 13:1–12. https://doi.org/10.1038/s41598-023-44961-3

FAO (2014) World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps

Hoang HTT, Van Rompaey A, Vu KC, et al (2020) An application of multilevel model for the analysis of factors influencing paddy field productivity in the Northern Vietnamese Mountains. Paddy Water Environ 18:153–166. https://doi.org/10.1007/s10333-019-00771-w

Hoffman L, Walters RW (2022) Catching Up on Multilevel Modeling. Annu Rev Psychol 73:659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Johnson PCD (2014) Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods Ecol Evol 5:944–946. https://doi.org/10.1111/2041-210X.12225

Kutka FJ, Smith ME (2007) How many parents give the highest yield in predicted synthetic and composite populations of maize? Crop Sci 47:1905–1913. https://doi.org/10.2135/cropsci2006.12.0802sc

Li Z, Taylor J, Yang H, et al (2020) A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. F Crop Res 248:107711. https://doi.org/10.1016/j.fcr.2019.107711

Liu Y, Heuvelink GBM, Bai Z, et al (2021) Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression. F Crop Res 264:108098. https://doi.org/10.1016/j.fcr.2021.108098

Masoni A, Calamai A, Marini L, et al (2020) Constitution of Composite Cross Maize (Zea mays L.) Populations Selected for the Semi-Arid Environment of South Madagascar. Agronomy 10:. https://doi.org/10.3390/agronomy10010054

Mukri G, Patil MS, Motagi BN, et al (2022) Genetic variability, combining ability and molecular diversity-based parental line selection for heterosis breeding in field corn (Zea mays L.). Mol Biol Rep 49:4517–4524. https://doi.org/10.1007/s11033-022-07295-3

Pedro C, Marçola MA, Charimba AM, et al (2023) Genetic potential of maize full-sib progenies subjected to a reciprocal recurrent selection. Pesqui Agropecu Bras 58:1–11. https://doi.org/10.1590/S1678-3921.pab2023.v58.03134

Reis EF dos, Reis MS, Cruz CD, Sediyama T (2004) Comparação de procedimentos de seleção para produção de grãos em populações de soja. Ciência Rural 34:685–692. https://doi.org/10.1590/s0103-84782004000300006

Wang H, Xu C, Liu X, et al (2017) Development of a multiple-hybrid population for genome-wide association studies: Theoretical consideration and genetic mapping of flowering traits in maize. Sci Rep 7:1–16. https://doi.org/10.1038/srep40239

Wang P, Zhang H, lyle D, et al (2019) Bulk pollen pollination in maize for efficient construction of introgression populations with high genome coverage. Plant Breed 138:252–258. https://doi.org/10.1111/pbr.12684

Yong H, Zhang F, Tang J, et al (2019) Breeding potential of inbred lines derived from five maize (Zea mays L .) populations. Euphytica 215:1–12. https://doi.org/10.1007/s10681-018-2319-8

Zhu W, Zhu S, Sunguya BF, Huang J (2021) Urban–rural disparities in the magnitude and determinants of stunting among children under five in tanzania: Based on tanzania demographic and health surveys 1991–2016. Int J Environ Res Public Health 18:. https://doi.org/10.3390/ijerph18105184

ARTIGO 4- high-throughput phenotyping of maize ear traits for yield prediction using tree-based machine learning and selection indexes

Periódico: Artificial Intelligence in Agriculture (Elsivier), versão preliminar.

César Pedro [a] *(https://orcid.org/0000-0002-2963-8652) and João Cândido de Souza [a]* (https://orcid.org/0000-0001-9580-4631)

[a] *Universidade Federal de Lavras, Departamento de Biologia, Aquenta Sol, Lavras - MG, CEP 37200-900, Brasil.*

*e-mail: cansouza@ufla.br (autor correspondente)

ABSTRACT

This study aimed to integrate high-throughput phenotyping, machine learning (ML) models, and selection indices to predict yield and optimize simultaneous genetic gains for the most important traits by selecting high-yielding multi-trait maize progenies. The experiment was conducted in 2024 at Muquém Farm (UFLA, Lavras-MG, Brazil), using an alpha-lattice design with 60 interpopulation progenies derived from crosses between populations A and B. Three main approaches were employed: (i) ear phenotyping through image analysis, (ii) ML models (Random Forest and Cubist) and OLS for comparison, and (iii) the MGIDI (Multi-trait Genotype-Ideotype Distance Index) selection index, constructed based on the most relevant traits identified by the models. The results demonstrated that all models were statistically significant, with Cubist showing the best predictive performance. The most influential traits for yield included Width and DFF (OLS), Width and DFM (Cubist), and Area, DFM, PH, and Width (RF). Although Cubist and RF models showed greater individual genetic gain, integrating Cubist and OLS with MGIDI maximized the selection response. It was concluded that the proposed strategy, combining phenotyping, ML, and MGIDI, is effective for predicting and selecting maize progenies, optimizing genetic gains. The methodology is replicable and adaptable to other crops, advancing precision breeding programs.

Keywords: *Zea mays L.*, Random Forest, Cubist, OLS, MGIDI.

Fenotipagem de alto rendimento de características de espigas de milho para predição da produtividade utilizando Aprendizado de Máquina baseado em árvores e índices de seleção

Resumo: Este estudo objetivou integrar fenotipagem de alto rendimento, modelos de aprendizado de máquina (ML) e índices de seleção para predizer a produtividade e otimizar os ganhos geneticos simultaneaos com as caracteristicas mais importantes por selecao de progênies multi-trait superiores de milho. O experimento foi conduzido em 2024 na Fazenda

Muquém (UFLA, Lavras-MG), utilizando um delineamento alfa-látice com 60 progênies interpopulacionais derivadas de cruzamentos entre as populações A e B. Foram empregadas três abordagens principais: (i) fenotipagem de espigas por análise de imagens, (ii) modelos de ML (Random Forest e Cubist) e OLS para comparação e (iii) o índice de seleção MGIDI (Multi-trait Genotype-ideotype Distance Index), construído com base em variáveis mais relevantes identificadas pelos modelos. Os resultados demonstraram que todos os modelos foram estatisticamente significativos, com o Cubist apresentando o melhor desempenho preditivo. As variáveis mais influentes na produtividade incluíram Width e DFF (OLS), Width e DFM (Cubist) e Area, DFM, PH e Width (RF). Embora os modelos Cubist e RF tenham mostrado maior ganho genético individual, a integração do cubist e OLS com o MGIDI maximizou a resposta de seleção. Conclui-se que a estratégia proposta, combinando fenotipagem, ML e MGIDI, é eficaz para a predição e seleção de progênies de milho, otimizando ganhos genéticos. A metodologia é replicável e adaptável a outras culturas agrícolas, oferecendo um avanço para programas de melhoramento de precisão.

Palavras-chaves: *Zea mays* L, Random forest, cubist, OLS, MGIDI.

1. Introduction

The pursuit of higher maize yield has benefited from advances in high-throughput phenotyping, machine learning (ML), and selection indices. Digital phenotyping enables automated, precise capture of large volumes of morphological traits that are difficult to measure manually, as well as rapid evaluation of multiple genotypes, enhancing efficiency in maize breeding (Makanza et al. 2018; Liang et al. 2021; Resende et al. 2024). However, using all measured traits can lead to time consuming analyses and reduce selection accuracy due to potential redundancies.

ML models are highly effective at handling high-dimensional data and identifying traits with the strongest predictive power for yield, improving prediction efficiency and accuracy (Babaie Sarijaloo et al. 2021; Dhaliwal and Williams 2024; Prasath et al. 2023). Yet, their standalone application faces challenges, such as the need for technical expertise and difficulties in translating results into practical breeding decisions.

Multi-trait selection indices have been widely used to optimize the selection of high-performance progenies by combining multiple morphological traits (Cruz et al. 2014; Olivoto and Nardino 2021). However, their effectiveness depends on including relevant traits, underscoring the importance of prior trait screening and analysis.

Thus, this study proposes integrating three approaches for predicting maize progeny yield: (i) Phenotyping maize ears via image analysis, (ii) yield prediction using decision tree-based ML models, and (iii) ML-guided multi-trait selection indices.

The synergy of these methodologies offers an innovative approach, leveraging the strengths of each model to optimize simultaneous trait selection and maximize genetic gains in maize breeding programs. This integrated strategy reduces selection redundancies and improves precision in identifying high-performance multi-trait progenies, ensuring more efficient and accurate outcomes.

2. Material and methods

From November 2023 to February 2024, two blocks (10 × 10 m each) of populations A and B were established, spaced two meters apart, to generate interpopulation AB half-sib progenies using the Bulk Pollen Pollination method. This involved mixing a random pollen sample from 10 plants to pollinate the upper ear of a plant from population B, resulting in 60 interpopulation half-sib progenies From March 2024 to July 2024, the progenies were evaluated in an alpha lattice design (10 × 6) with two replications at the experimental field of Muquém Farm (21°12'06.7"S, 44°58'45.2"W, elevation: 918.84 m), located at the Center for Scientific and Technological Development in Agriculture at the Federal University of Lavras, in southern Minas Gerais State, Brazil.

The experimental area has a Latossolo (Oxisol) soil type, according to Santos et al. (2018). The climate is classified as humid temperate (Cwa), characterized by dry winters and rainy summers. During the experimental period, the mean temperature and precipitation in the area were 22.96 °C and 218.4 mm, respectively (data from Lavras weather station, code: 83687; geographic coordinates: 21°13'34.0"S, 44°58'47.0"W).

The seeding rate was four seeds per linear meter in four-meter-long plots spaced 0.6 meters apart. Base fertilization was applied at 250 kg ha^{-1} of NPK 8-28-16, and topdressing was performed 25 days after planting using 200 kg ha^{-1} of urea (45% N). Other crop management practices followed regional recommendations (Borém, 2017). The evaluated traits included grain yield (GY, ton ha^{-1}), days to male flowering (DMF), days to female flowering (DFF), and plant height (PH, cm).

2.1. Maize ear phenotyping

The evaluated progenies were submitted for image capture, with three ears per replication. A custom wooden box setup was used, featuring overhead artificial lighting, a contrasting blue background, and a professional Canon EOS 60D camera with a 35 mm lens positioned at the top center of the box. The camera was connected to a computer to save RGB images in JPG format (Fig. 1A and B). The process began with the pre-processing of the saved images, removing unnecessary margins using ImageJ software (Ferreira and Rasband, 2012) (Fig. 1C). Next, image binarization was performed to separate the ears from the blue background using spectral bands, including BGI, HUE2, CI, and NR. These bands were used to determine the contrast required for isolating the ears. Subsequently, Otsu's thresholding method was applied, with CI adjustment (Fig. 1D).

During the segmentation stage, the images were divided into two regions using the CI and BGI bands. This segmentation was fine-tuned with inversion to ensure proper separation of the ears from other image areas. Following this, segmented ears were identified. After identification, ear measurements were taken, extracting multiple traits; however, only length, width, and area were included in the models (Fig. 1F). Pixel measurements were converted to centimeters using the image's pixel density (DPI). This conversion allowed the results to be interpreted in real-world units. Image processing and analysis were conducted in R (Core Team, 2024) using the Pliman package (Olivoto, 2022).

To validate phenotyping, manual measurements of length and width traits were taken using a conventional ruler and caliper, respectively. Subsequently, genetic variance, heritability, and accuracy were estimated for all traits. Pearson's correlation and mean absolute error between manually obtained traits and image-based analysis were also calculated.

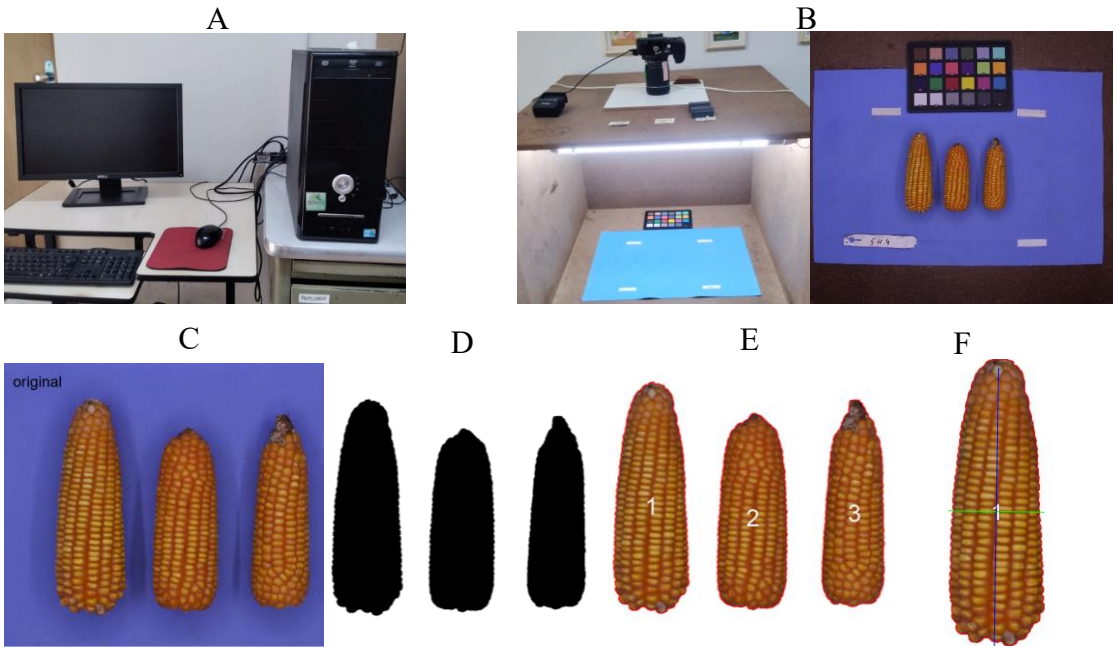


Fig. 1. Computer (A), Original unprocessed image (B), Preprocessed original image (C), Segmented image (D), Ear identification in the image (E), Measurement of ear length (blue line), ear width (green line), and other corn ear traits (F).

2.2 Tree-based machine learning models

Analysis began with the construction of three ML models to predict grain yield (GY). The developed ML models included Random Forest (RF), Cubist, and Ordinary Least Square Regression (OLS) model for comparison purposes. Cubist is a predictive regression model that combines concepts from Quinlan (1992, 1993). It starts with a tree structure, converting each path into a rule. For each rule, a regression model is fitted based on the corresponding data. Rules are pruned or merged, and the variables used in the pruned sections become predictors for the linear regression models. This process aligns with the "M5" or "Model Tree" approach (Max et al., 2024).

Next, validation was performed using Leave-One-Out Cross-Validation (LOOCV) via the caret package (Max et al., 2024), where a single data point served as the test set while the remaining points formed the training set. This process was repeated for each data point, ensuring robust validation despite the small dataset size. After validation, the test data were

stored to predict new out-of-training-sample data. The observed and predicted values were calculated to evaluate model performance using Pearson correlation and root mean square error (RMSE).

In the next step, the decision trees for the models were constructed. The trees were generated using the iml package (Molnar, 2018), which provides tools for the interpretability of ML models. Additionally, Principal Component Analysis (PCA) was applied to visualize the clustering (nodes) of the progenies resulting from the decision trees.

2.3 Multi-trait selection indices

Subsequently, the most relevant traits for splitting the decision trees of each model were identified. These traits were used to construct the Multi-Trait Genotype–Ideotype Distance Index (MGIDI), as proposed by Olivoto and Nardino (2021). This index was employed to identify high-performance mult-trait progenies and estimate selection gains based on the desired ideotype. The traits associated with the ideotype included: Positive selection direction for grain yield (GY), ear width (E_width), ear length (E_length), and ear area (E_area). Negative selection direction for days to female flowering (DFF), days to male flowering (DMF) and plant height (PH).

Progeny selection was performed individually for each model (Random Forest, Cubist, and OLS) and also in a combined approach using the MGIDI index with a selection intensity of 15%. The results were visualized in a Venn diagram, allowing for the assessment of coincidence rates among selected progenies across the different models. This final step was essential to validate the consistency of the ML models in progeny selection, as well as to test their effectiveness and reliability in selection.

3. Results and discussion

Table 1 presents genetic and phenotypic parameters for maize traits measured manually and digitally. Significant genetic variability was detected for all traits, with moderate to high heritability and high selection accuracy, indicating strong breeding potential. Digital measurements generally exhibited lower absolute error; however, ear width had a smaller error

than ear length. Genetic parameters and means were consistent across measurement methods, supporting the efficiency of digital phenotyping. The high heritability and accuracy observed reinforce the reliability of these methods, aligning with findings from Yang et al. (2021) who highlight the importance of high-throughput phenotyping for selecting complex traits.

A strong positive and significant correlation was observed between the two measurement methods (Fig. 2A). However, the correlation was higher for ear width associated with lower absolute error than for ear length (Table 1, Fig. 2A), which may be attributed to potential inaccuracies in manual length measurements. The precision of image-based data was discussed by Makanza et al. (2018) and Liang et al. (2021), who note that imaging technologies can be as accurate as manual measurements when properly calibrated and adapted to field conditions. Nevertheless, these authors emphasize that precision may vary depending on the trait being measured.

Table 1

Phenotypic and genetic parameters of the traits (Grain yield: GY, Days to male flowering: DMF, Days to female flowering: DFF, Plant height: PH, Ear width: E_width, Ear length: E_length, Ear area: E_area) in maize progenies.

Traits	$\hat{\sigma}^2_{g1}$ ($\hat{\sigma}^2_{g2}$)	h^2_1 (h^2_2)	Ac1 (Ac2)	u1(u2)	MAE
GY (t ha^{-1})	2.19 ***	0.77	0.89	6.75	–
E_area (cm^2)	13.10 ***	0.58	0.76	44.64	–
E_width (cm)	0.045 (0.036)***	0.57 (0.55)	0.75 (0.74)	4.20 (4.11)	0.15
E_length (cm)	0.68 (0.74)***	0.54 (0.61)	0.74 (0.78)	14.79 (13.59)	1.42
DMF (days)	4.70 ***	0.84	0.92	63.79	–
DFF (days)	5.49 ***	0.84	0.917	63.46	–
PH (cm)	342 ***	0.69	0.83	161.47	–

$\hat{\sigma}^2_{g1}, h^2_1$, Ac1 and u1= Genetic variance, heritability, accuracy, and mean of manual measurements. $\hat{\sigma}^2_{g2}, h^2_2$, Ac2 and u2= Genetic variance, heritability, accuracy, and mean of digital measurements. MAE: Mean Absolute Error. *** Significant at 0.001 by the Likelihood Ratio Test. Values outside parentheses are manual measurements (both manual and digital measurements were possible for E_width and E_length), while values inside parentheses are digital measurements.

According to Makanza et al. (2018) and Resende et al. (2024), image-based phenotyping accelerates breeding programs by enabling rapid and efficient data collection. This is particularly relevant for crops like maize, where large numbers of progenies and traits must be evaluated. Manual phenotyping, on the other hand, can introduce bias and variability due to differences between observers or even within a single operator. Duddu et al. (2019) state that standardizing measurements obtained through imaging reduces human bias, resulting in more consistent and comparable data. Thus, image-based standardization provides a significant

advantage in obtaining uniform data by eliminating variability introduced by multiple operators or even a single operator.

Fig. 2B illustrates the relationship between observed and predicted grain yield in maize progenies using Cubist, Random Forest (RF), and Ordinary Least Squares (OLS) models. All three models showed a significant correlation between observed and predicted yield values. The Cubist model outperformed the RF model, with both machine learning approaches proving more effective than OLS. These results indicate that Cubist was more efficient in capturing grain yield variability among maize progenies. For maize breeding programs, the ability to predict grain yield with high accuracy is critical, particularly when identifying high-performance progenies efficiently. The high performance of Cubist suggests it could be a valuable tool for predicting and selecting progenies in maize breeding, leading to a faster and more precise selection process.

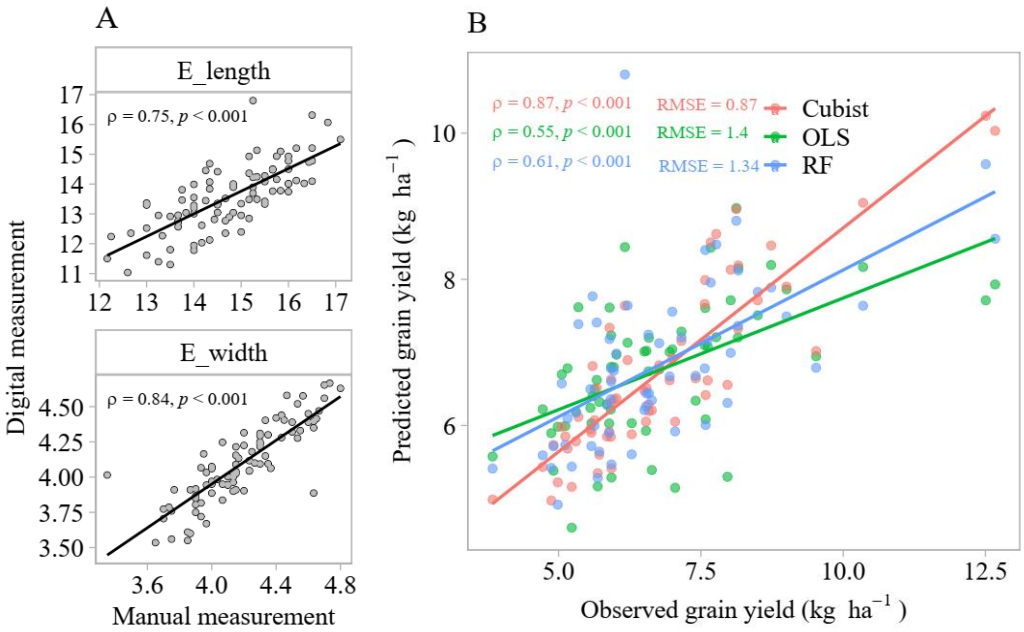


Fig. 2. (A) Relationship between manual and digital measurements. (B) Relationship between observed and predicted yield based on Random Forest (RF), Cubist, and Ordinary Least Squares (OLS) in maize progenies.

Fig. 3 presents the results of tree-based ML models and OLS, along with their interpretability for selection purposes in maize breeding. These models highlight the most important predictors influencing yield. The OLS model emphasized E_width and DFF, while the Cubist model highlighted E_width and DFM. In contrast, the RF model identified E-area,

DFM, E_width, and PH as the most significant predictors of grain yield (Figure 3: A, C, and E). The OLS and Cubist models were simpler than the RF model, which exhibited greater complexity in terms of the relationship between observed and predicted yield.

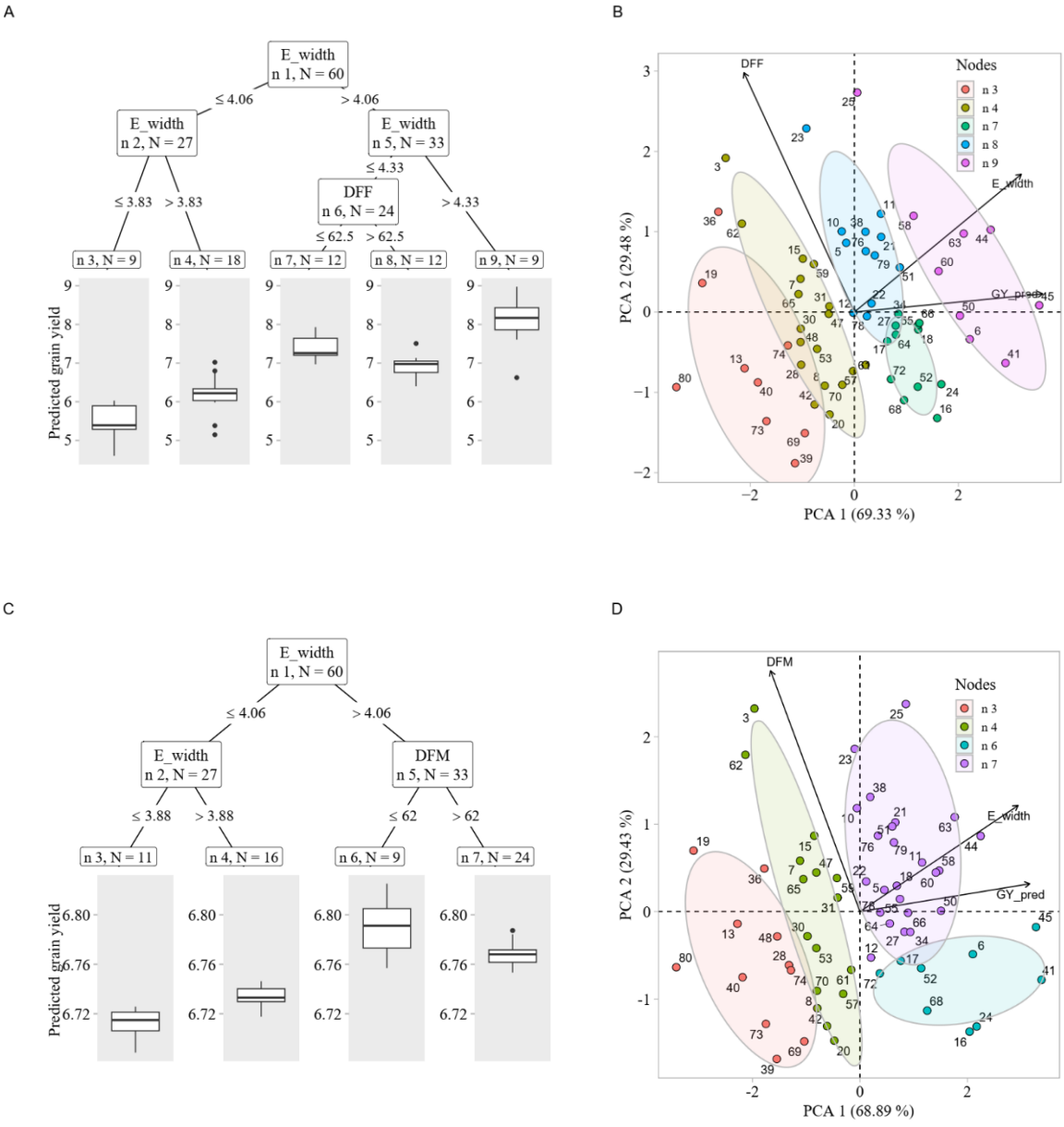
The findings of this study align with those of Khanal et al. (2021) and Chen et al. (2021) who applied various ML models and identified the Cubist model as the most effective for predicting maize yield, followed by the RF model. However, it is worth noting that Babaie Sarijaloo et al. (2021) and Prasath et al. (2023) reached different conclusions, suggesting that RF was the most suitable model for this purpose. This discrepancy may be attributed to the fact that the latter studies did not include the Cubist model in their analyses, potentially limiting a direct comparison between methods. Therefore, the superiority of the Cubist model observed in our study, as well as in the works of Khanal et al. (2021) and Chen et al. (2021), suggests that the Cubist model was essential for a more accurate assessment of maize yield.

The relevance of predictors such as PH (Reis et al., 2009), E_area, E_width (Makanza et al., 2018; Resende et al., 2014), and DFF (Amaral et al. 2023) for yield prediction aligns with the literature, which recognizes morpho-agronomic traits as key traits for predicting and selecting maize progenies. The RF model captured E_area, whereas the OLS and Cubist models emphasized E_width as the most influential predictors of yield. This suggests that ML and OLS models identify distinct patterns in agronomic traits as primary yield predictors, reflecting different approaches in capturing complex trait interactions. These differences are significant, as they indicate that a combined use of these models could provide a more comprehensive understanding of the key traits affecting yield.

Fig. 3 (B, D, and F) demonstrate the practical interpretability of the models by displaying the predicted progenies at the terminal nodes of the decision trees. Principal component analysis (PCA) further enabled the visualization of progeny-trait relationships and highlighted the phenotypic divergence among different progeny groups at the terminal nodes. This approach facilitates the identification and selection of high-performance and divergent progenies, offering strategic insights for decision-making in breeding programs.

Studies by Belalia et al. (2019) and Leite et al. (2019) emphasize the importance of PCA for dimensionality reduction in plant breeding, enabling more targeted selection of high-yielding progenies. Research also indicates that interpretable ML models enhance researchers' confidence in applying these tools to agriculture and plant breeding, where precise and

explainable predictions are critical for decision-making (Blanco-justicia et al. 2020; Jones et al. 2022).



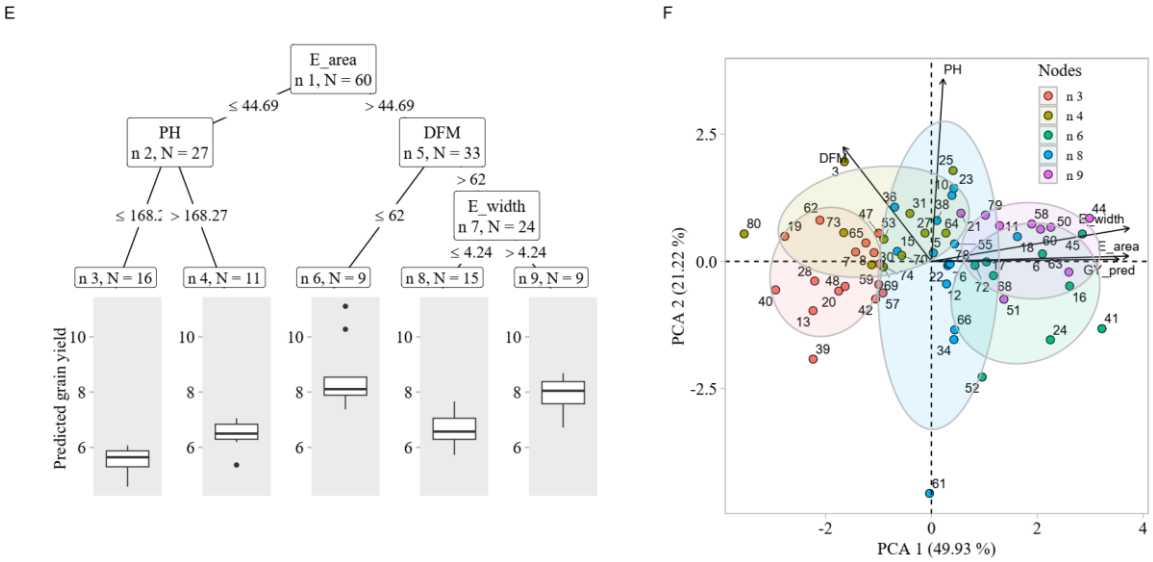


Fig. 3. Predicted yield (t ha^{-1}) of maize progenies using tree-based Machine Learning models: Cubist (C and D), Random Forest (RF) (E and F), and Ordinary Least Squares (OLS) (A and B). Decision Trees (A, C, and E) and graphical visualization (B, D, and F). Principal Component Analysis (PCA).

Table 2 presents the impact of the ML and OLS models individually compared to MGIDI, as well as the combination of these models with the MGIDI index on simultaneous selection gains expressed as a percentage for key yield predictors and grain yield in maize.

For each model, the two best progeny nodes with the highest predicted performance (Figure 3) were selected, and genetic gains were estimated for both. While the RF model was not the best overall, it demonstrated superior genetic gain prediction compared to OLS, Cubist, and MGIDI in the two best-selected nodes. However, in the node with the highest predicted mean yield, RF was similar to Cubist, whereas the gains from OLS were comparable to those from MGIDI.

The genetic gains resulting from the combination of these models with MGIDI surpassed those obtained with the individual models, highlighting the importance of integration. Although the RF+MGIDI combination showed considerable gains, the OLS+MGIDI and Cubist+MGIDI combinations delivered the highest genetic gains by identifying high-performance progeny with desirable traits such as E-width, DFF, and DMF. Nevertheless, the similarity in gains between these two models suggests that DFF and DMF may be interchangeable.

Based on the results, maximizing predicted gains depends more on the traits prioritized by the ML models for constructing selection indices, and the model with the highest predictive power does not always lead to the greatest genetic gain when integrated with a selection index. Individually, the Cubist model performed best, followed by RF and OLS in predictive ability. However, when combined with selection indices, Cubist and OLS showed higher genetic gain responses than RF. It is worth noting that, despite differences in predictive power, all models exhibited significant predictive performance, which is essential for the optimal selection of high-performance progeny.

Therefore, it is evident that when the MGIDI index was applied considering all traits, the selection gains were lower, reinforcing the importance of using only the most relevant traits, those identified by the ML models. This approach not only improves the accuracy of selection gains but also simplifies the process by reducing 'noise' from irrelevant traits and maximizing genetic gains.

Studies such as those by Chlingaryan et al. (2018) and Pham et al. (2022) suggest that using the most important traits increases the predictive accuracy of models, promoting a reduction in the number of traits and making the selection process more targeted and resource-efficient. Thus, integrating machine learning models with selection indices, while utilizing only the most significant variables, provides an efficient and focused method for predicting and selecting high-performance progenies in the breeding program.

Table 3

Estimated heritability and percentage gains for traits from selecting the top 15% of maize progenies using decision tree-based models (Cubist and Random Forest) and OLS, integrated by the Multi-Trait Genotype-Ideotype Distance Index (MGIDI).

Modelo	Trait	FA	X_o	X_s pred	X_s obs	SG %
	GY	_	6.75	7.4 - 8.05	7.80 - 7.94	11.96 - 13.56
OLS	E_width	_	4.09	_	4.18 - 4.46	1.50 - 6.16
	DFF	_	63.50	_	61.21 - 62.73	-3.03 - -1.02
	GY	_	6.75	6.79 - 6.77	8.70 - 7.04	22.22 - 3.30
Cubist	E_width	_	4.09	_	4.26 - 4.27	2.83 - 3.00
	DMF	_	63.80	_	60.68 - 64.37	-4.12 - 0.75
	GY	_	6.75	8.55 - 7.92	8.70 - 8.29	22.22 - 17.54
	E-area	_	44.60	_	47.83 - 50.15	4.21 - 7.23
RF	E_width	_	4.09	_	4.26 - 4.37	2.83 - 4.66
	DMF	_	63.80	_	60.68 - 64.16	-4.12 - - 0.48
	PH	_	161.00	_	167. 00 - 181. 90	2.55 - 8.89
	GY	FA1	6.75	8.70	9.26	28.80
OLS+MGIDI	E_width	FA1	4.09	4.30	4.40	5.12
	DFF	FA1	63.50	61.40	61.10	-3.17
	GY	FA1	6.75	8.70	9.26	28.80
Cubist+MGIDI	E_width	FA1	4.09	4.30	4.40	5.10
	DFM	FA1	63.80	61.90	61.60	-2.94
	GY	FA1	6.75	8.36	8.83	23.8
	E_area	FA1	44.60	47.40	49.30	6.09
RF+MGIDI	E_width	FA1	4.09	4.24	4.29	3.71
	DMF	FA2	63.80	62.10	61.80	-2.73
	PH	FA2	161.00	152.00	146.00	-6.08
	GY	FA1	6.75	7.65	7.95	13.40
	E_area	FA1	44.60	48.00	50.70	7.57
	E_width	FA1	4.09	4.25	4.33	3.94
MGIDI	E_lenght	FA1	13.40	14.10	14.60	5.25
	DMF	FA2	63.80	62.20	61.90	-2.44
	DFF	FA2	63.80	61.50	61.20	-3.03
	PH	FA3	161.00	161.00	161.00	0.00

-: Indicator of variation range, _ no data. FA: Factor. Xo and Xs: Overall mean and selected mean, respectively. Obs and Pred: Observed and predicted yield, respectively. SG%: Percentage selection gain.

Figure 4 shows the overlap of the top-performing progenies identified in the two best nodes of each model, as well as the combination of the models with the MGIDI index. The results indicate that approximately 35% of the progenies were consistently selected by the RF model and its combination with MGIDI. On the other hand, the OLS and Cubist models

captured all progenies from their combinations with MGIDI, with an overlap of 42.9% and 50%, respectively. This suggests that the OLS and Cubist models, when used independently, were as effective in predicting high-performance progenies as their combinations with MGIDI, significantly outperforming the simultaneous selection of all traits by MGIDI alone.

Additionally, these models provided insights into the formation of divergent groups, which can be recombined to enhance genetic variability in the breeding program. The findings suggest that incorporating ML models into breeding strategies, whether used alone or in combination, optimizes selection efficiency and genetic gains in maize improvement. Recent studies emphasize the importance of ML in predicting key traits and generating valuable insights (Babaie Sarijaloo et al. 2021; Prasath et al., 2023).

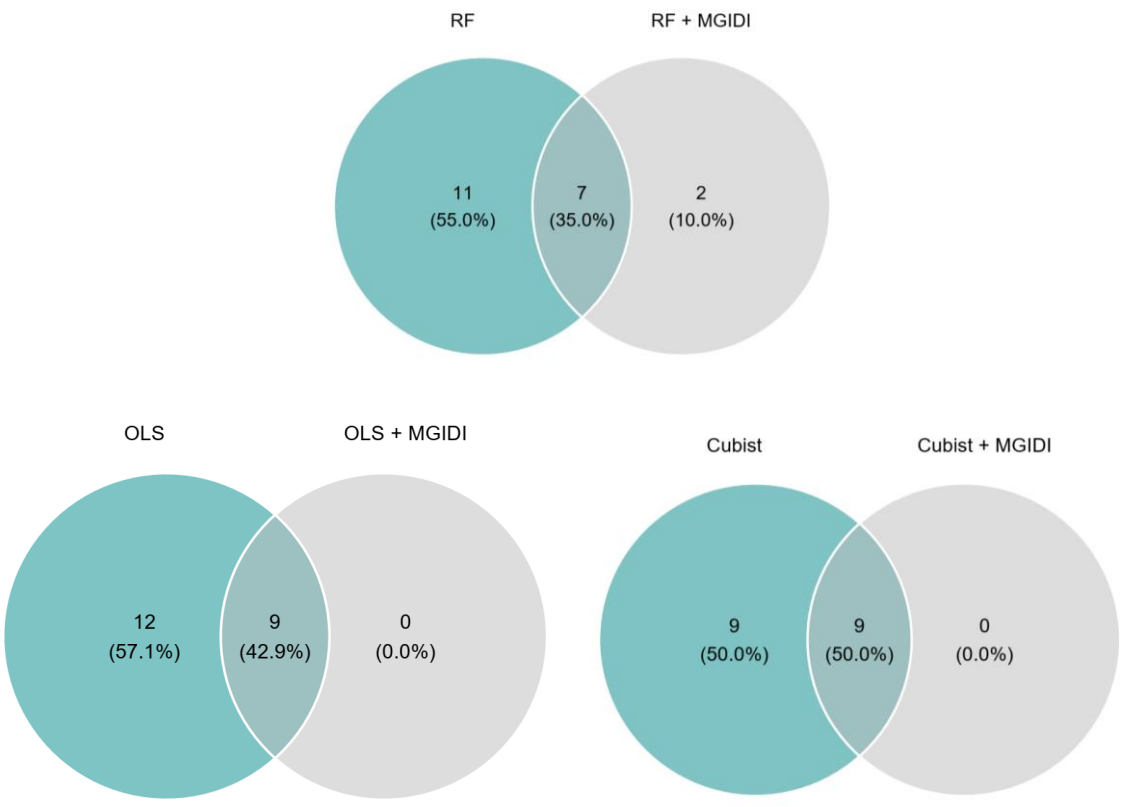


Fig. 4. Progeny selection coincidence index based on Decision Trees from Machine Learning models (Cubist and RF) and OLS, and its integration with the Multi-Trait Genotype-Ideotype Distance Index (MGIDI).

4. Conclusion

The study demonstrated the effectiveness of integrating high-throughput phenotyping with machine learning models. The most influential Ear width and Ear area. The Cubist and Random forest models showed higher individual genetic gain. The integration of the Cubist and Ordinary Least Squares models with the Multi-Trait Genotype–Ideotype Distance Index optimized the response to simultaneous trait selection by identifying the best progenies.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

To Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, finance code 001), Brazil, for supporting this research.

References

Amaral, L. de O., Miranda, G.V., Souza, J. da S., Moitinho, A.C.R., Cristeli, D.S., Silva, H.K. da, Anjos, R.S.R. dos, Alliprandini, L.F., Unêda-Trevisoli, S.H., 2023. Application of Artificial Neural Networks to Predict Genotypic Values of Soybean Derived from Wide and Restricted Crosses for Relative Maturity Groups. Agronomy 13, 1–12. https://doi.org/10.3390/agronomy13102476

Babaie Sarijaloo, F., Porta, M., Taslimi, B., Pardalos, P.M., 2021. Yield performance estimation of corn hybrids using machine learning algorithms. Artif. Intell. Agric. 5, 82–89. https://doi.org/10.1016/j.aiia.2021.05.001

Belalia, N., Lupini, A., Djemel, A., Morsli, A., Mauceri, A., Lotti, C., Khelifi-Slaoui, M., Khelifi, L., Sunseri, F., 2019. Analysis of genetic diversity and population structure in Saharan maize (Zea mays L.) populations using phenotypic traits and SSR markers. Genet. Resour. Crop Evol. 66, 243–257. https://doi.org/10.1007/s10722-018-0709-3

Blanco-justicia, A., Domingo-ferrer, J., Blanco-justicia, A., Machine, J.D., Explainability, L., 2020. Machine Learning Explainability Through Comprehensible Decision Trees To cite this version : HAL Id : hal-02520062 Comprehensible Decision Trees 0–12.

BORÉM, A; GALVÃO, J. C. C.; PIMENTEL, M.A., 2017. No TitleMilho: do plantio à colheita., 2nd ed. Viçosa.

Chen, X., Feng, L., Yao, R., Wu, X., Sun, J., Gong, W., 2021. Prediction of maize yield at the city level in China using multi-source data. Remote Sens. 13, 1–17. https://doi.org/10.3390/rs13010146

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield

prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Cruz, Cosme Damião; Carneiro, Pedro Crescencio souza; Regazzi, A.J., 2014. Modelos biometricos aplicados ao melhoramento genetico, 3rd ed. 2014, Vicosa.

Dhaliwal, D.S., Williams, M.M., 2024. Sweet corn yield prediction using machine learning models and field-level data. Precis. Agric. 25, 51–64. https://doi.org/10.1007/s11119-023-10057-1

Duddu, H.S.N., Johnson, E.N., Willenborg, C.J., Shirtliffe, S.J., 2019. High-throughput UAV image-based method is more precise than manual rating of herbicide tolerance. Plant Phenomics 2019, 6036453. https://doi.org/10.34133/2019/6036453

Ferreira, T., Rasband, W., 2012. ImageJ User Guide User Guide ImageJ. Image J user Guid. 1.46r. https://doi.org/10.1038/nmeth.2019

Humberto Gonçalves dos Santos, Paulo Klinger Tito Jacomine, Lúcia Helena Cunha dos Anjos, Virlei Álvaro de Oliveira, José Francisco Lumbreras, Maurício Rizzato Coelho, Jaime Antonio de Almeida, José Coelho de Araújo Filho, João Bertoldo de Oliveira, T.J.F.C., 2018. No TitleSistema Brasileiro de Classificação de Solos, 5th ed. Empresa Brasileira de Pesquisa Agropecuária Embrapa Solos, Embrapa Brasília, DF.

Jones, E.J., Bishop, T.F.A., Malone, B.P., Hulme, P.J., Whelan, B.M., Filippi, P., 2022. Identifying causes of crop yield variability with interpretive machine learning. Comput. Electron. Agric. 192, 106632. https://doi.org/10.1016/j.compag.2021.106632

Khanal, S., Klopfenstein, A., KC, K., Ramarao, V., Fulton, J., Douridas, N., Shearer, S.A., 2021. Assessing the impact of agricultural field traffic on corn grain yield using remote sensing and machine learning. Soil Tillage Res. 208, 104880. https://doi.org/10.1016/j.still.2020.104880

Leite, P.H.M.P., Silva, V.P. da, Gilio, T.A.S., Azevedo, R.F., Oliveira, T.C. de, Barelli, M.A.A., 2019. Diversidade genética em cultivares e linhagens de feijão comum (Phaseolus vulgaris L.) utilizando análises multivariadas. Cult. Agronômica Rev. Ciências Agronômicas 28, 268–279. https://doi.org/10.32929/2446-8355.2019v28n3p268-279

Liang, X., Ye, J., Li, X., Tang, Z., Zhang, X., Li, W., Yan, J., Yang, W., 2021. A high-throughput and low-cost maize ear traits scorer. Mol. Breed. 41. https://doi.org/10.1007/s11032-021-01205-4

Makanza, R., Zaman-Allah, M., Cairns, J.E., Eyre, J., Burgueño, J., Pacheco, Á., Diepenbrock, C., Magorokosho, C., Tarekegne, A., Olsen, M., Prasanna, B.M., 2018. High-throughput method for ear phenotyping and kernel weight estimation in maize using ear digital imaging. Plant Methods 14, 1–13. https://doi.org/10.1186/s13007-018-0317-4

Max, A., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Ziem, A., Scrucca, L., Hunt, T., Kuhn, M.M., 2024. Package ‘ caret .’

Molnar, C., 2018. iml: An R package for Interpretable Machine Learning. J. Open Source Softw. 3, 786. https://doi.org/10.21105/joss.00786

Olivoto, T., 2022. Lights, camera, pliman! An R package for plant image analysis, Methods in Ecology and Evolution. https://doi.org/10.1111/2041-210X.13803

Olivoto, T., Nardino, M., 2021. Genetics and population analysis MGIDI : toward an effective multivariate selection in biological experiments 37, 1383–1389. https://doi.org/10.1093/bioinformatics/btaa981

Pham, H.T., Awange, J., Kuhn, M., 2022. Evaluation of Three Feature Dimension Reduction Techniques for Machine Learning-Based Crop Yield Prediction Models. Sensors 22, 1–18. https://doi.org/10.3390/s22176609

Prasath, N., Sreemathy, J., Krishnaraj, N., Vigneshwaran, P., 2023. Analysis of Crop Yield Prediction Using Random Forest Regression Model. Smart Innov. Syst. Technol. 324, 239–249. https://doi.org/10.1007/978-981-19-7447-2_22

Resende, E.L., Bruzi, A.T., Cardoso, E. da S., Carneiro, V.Q., Pereira de Souza, V.A., Frois Correa Barros, P.H., Pereira, R.R., 2024. High-Throughput Phenotyping: Application in Maize Breeding. AgriEngineering 6, 1078–1092. https://doi.org/10.3390/agriengineering6020062

Yang, S., Zheng, L., He, P., Wu, T., Sun, S., Wang, M., 2021. High-throughput soybean seeds phenotyping with convolutional neural networks and transfer learning. Plant Methods 17, 1–17. https://doi.org/10.1186/s13007-021-00749-y

TERCEIRA PARTE

Considerações gerais

O estudo abordou a aplicação de metodologias estatísticas e de Machine Learning (ML) para otimizar a Seleção Recorrente Recíproca (SRR), visando à precisão e à eficiência na escolha das progênies de milho promissoras, capazes de maximizar os ganhos genéticos. A quantificação do potencial genético das populações, as análises multicaracterísticas, o uso integrado de fenotipagem de alto rendimento, Machine Learning e índices de seleção, além da abordagem de análise multinível, foram estratégias cruciais para a otimização da seleção e a obtenção precisa de ganhos genéticos. O estudo valida o uso do ML como uma ferramenta alternativa de grande relevância no programa de SRR. Essa pesquisa pode ser replicada em diferentes contextos de estudo do milho e de outras culturas agrícolas.