

BRENDA VIEIRA DE OLIVEIRA

**AVALIAÇÃO DO IMPACTO DA SELEÇÃO BASEADA NA COMBINAÇÃO DE
ANÁLISE DE FATORES E ÍNDICES DE SELEÇÃO EM UM PROGRAMA DE
MELHORAMENTO DE SELEÇÃO RECORRENTE**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Camila Ferreira Azevedo

Coorientadores: Ana Carolina Campana
Nascimento
Leísa Pires Lima
Moysés Nascimento

VIÇOSA - MINAS GERAIS

2024

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

T

Oliveira, Brenda Vieira de, 2000-
O48a Avaliação do impacto da seleção baseada na combinação de
2024 análise de fatores e índices de seleção em um programa de
 melhoramento de seleção recorrente / Brenda Vieira de Oliveira.
 – Viçosa, MG, 2024.
 1 dissertação eletrônica (58 f.): il. (algumas color.).

Orientador: Camila Ferreira Azevedo.
Dissertação (mestrado) - Universidade Federal de Viçosa,
Departamento de Estatística, 2024.
Referências bibliográficas: f. 54-58.
DOI: <https://doi.org/10.47328/ufvbbt.2024.699>
Modo de acesso: World Wide Web.

1. Análise multivariada. 2. Melhoramento genético -
Métodos estatísticos. I. Azevedo, Camila Ferreira, 1988-.
II. Universidade Federal de Viçosa. Departamento de Estatística.
Programa de Pós-Graduação em Estatística Aplicada e
Biometria. III. Título.

CDD 22. ed. 519.5354

BRENDA VIEIRA DE OLIVEIRA

**AVALIAÇÃO DO IMPACTO DA SELEÇÃO BASEADA NA COMBINAÇÃO DE
ANÁLISE DE FATORES E ÍNDICES EM UM PROGRAMA DE MELHORAMENTO
DE SELEÇÃO RECORRENTE**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 28 de agosto de 2024.

Assentimento:

Documento assinado digitalmente
 **BRENDA VIEIRA DE OLIVEIRA**
Data: 15/10/2024 14:24:59-0300
Verifique em <https://validar.it.gov.br>

Brenda Vieira de Oliveira
Autor

Documento assinado digitalmente
 **CAMILA FERREIRA AZEVEDO**
Data: 15/10/2024 12:13:44-0300
Verifique em <https://validar.it.gov.br>

Camila Ferreira Azevedo
Orientador

AGRADECIMENTOS

A Deus, primeiramente, por todas as vezes em que pensei que não conseguiria e Ele me deu forças. Obrigada Senhor por Tua imensa misericórdia e proteção, por Teu imenso amor e amparo. Sem Ti, eu nada seria.

À minha mãe, Nilzângela, por ser meu exemplo de força e dedicação, por nunca medir esforços para que eu chegasse até aqui, e pelo carinho e amor incondicionais. Obrigada por me ensinar a ser forte e perseverante.

À minha orientadora, professora Camila Ferreira Azevedo, pela dedicação, disponibilidade e paciência, pelo suporte constante, pelos ensinamentos valiosos e conselhos que foram fundamentais para o meu crescimento profissional. Muito obrigada por tudo!

Aos meus irmãos, Breno e Bruna, por sempre acreditarem em mim, por me incentivarem a seguir em frente mesmo nos momentos mais difíceis, e por todo o amor e carinho que me deram ao longo desta jornada.

Ao meu pai, minha avó Carmelita, e minhas tias Marilene e Joelma, por fazerem tudo ao seu alcance para me ajudar e me apoiar. Vocês foram uma fonte constante de apoio e motivação, e sou imensamente grata por tê-los ao meu lado.

Ao meu namorado, Rossir Junior, por todo o amor, carinho, incentivo e apoio. Sua presença e motivação foram fundamentais para que eu superasse os desafios ao longo desta trajetória. Sou imensamente grata por ter você ao meu lado, acreditando em mim e me inspirando todos os dias.

À toda sua família também, Yêda, Rossir, Marisa, Leísa, Osvaldo, Miguel, Júlia, Rafael e Sillas, por me acolherem como uma segunda família, por todo carinho e pelo incentivo. Obrigada por estarem ao meu lado e por fazerem parte desta jornada.

À minhas amigas Mariana e Sara, que estiveram ao meu lado em todos os momentos até aqui, por todo apoio, torcida e pela nossa amizade. Obrigada por tornarem esta jornada mais leve e alegre!

Aos meus amigos do MatMigos e da salinha do PPESTBIO, pelos momentos maravilhosos ao longo desta jornada, por torcerem por mim e pela amizade que construímos. Sou imensamente grata a cada um de vocês.

Aos doutores e meus coorientadores Ana Carolina Campana Nascimento, Leísa Pires Lima e Moisés Nascimento por toda ajuda, disponibilidade e sugestões para melhoria deste trabalho.

A todos os professores do PPESTBIO por contribuírem com seus ensinamentos e conselhos em minha formação acadêmica.

Ao secretário do PPESTBIO, Junior, pelo encorajamento e ajuda contínuos, assim como às secretárias do DET, Lays e Nayara, por toda assistência.

À Universidade Federal de Viçosa, por todas as oportunidades acadêmicas até aqui, tanto na graduação quanto no mestrado. Sou imensamente grata a esta instituição que mudou completamente a minha vida, não só profissionalmente, mas também pessoalmente, proporcionando experiências enriquecedoras e aprendizados que levarei para sempre comigo.

À banca examinadora, Professor Filipe Ribeiro Formiga Teixeira e Professora Gabi Nunes Silva, pela disponibilidade e sugestões para o aperfeiçoamento deste trabalho.

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

E a todos aqueles que contribuíram de alguma forma, direta ou indiretamente, para a concretização deste trabalho, meus sinceros agradecimentos.

"A perseverança não é uma longa corrida; ela é muitas corridas curtas, uma após a outra." – Walter Elliott

RESUMO

OLIVEIRA, Brenda Vieira de, M.Sc. Universidade Federal de Viçosa, outubro de 2024.
Avaliação do impacto da seleção baseada na combinação de análise de fatores e índices de seleção em um programa de melhoramento de seleção recorrente.
Orientador: Camila Ferreira Azevedo. Coorientadores: Ana Carolina Campana Nascimento, Leísa Pires Lima e Moysés Nascimento.

Com a crescente demanda comercial, o desenvolvimento de novas cultivares pelos programas de melhoramento exige que os genótipos selecionados atendam a uma ampla gama de características, incluindo resistência a doenças, produtividade e qualidade, entre outras. O índice de seleção é uma técnica biométrica amplamente utilizada para selecionar genótipos superiores, baseando-se em várias características simultaneamente, mas enfrenta desafios em sua confiabilidade quando muitos atributos estão envolvidos, devido às correlações entre os caracteres e à estimativa dos pesos do índice. Para superar essa questão, uma técnica multivariada que visa reduzir a dimensão do conjunto de dados e garantir a ortogonalidade entre as variáveis latentes é a análise de fatores (AF). Assim, com o objetivo de avaliar a combinação das técnicas a longo prazo, foram aplicados os escores da AF em substituição às variáveis fenotípicas no índice, comparando essa abordagem ao uso exclusivo do índice. Diversos cenários foram simulados, com variações na quantidade de características em seleção e as direções e magnitudes das correlações entre elas, para avaliar o impacto na seleção. A eficiência da combinação foi avaliada em termos de ganho genético, diversidade genética e endogamia, sendo esses resultados comparados ao uso exclusivo do índice de seleção. Os resultados indicam que, a longo prazo, a combinação das técnicas apresentou ganhos genéticos semelhantes ao índice tradicional, mas com maior diversidade genética na maioria dos cenários. O primeiro capítulo deste trabalho é uma revisão bibliográfica dos métodos utilizados, enquanto o segundo capítulo descreve a execução do trabalho.

Palavras-chave: Simulação; Correlação; Ganho Genético; Diversidade Genética.

ABSTRACT

OLIVEIRA, B. V., M.Sc. Universidade Federal de Viçosa, October, 2024. **Assessment of the impact of selection based on the combination of factor analyzes and selection indices in a recurrent selection breeding program.** Adviser: Camila Ferreira Azevedo. Co-advisers: Ana Carolina Campana Nascimento, Leísa Pires Lima e Moysés Nascimento.

With increasing commercial demands, the development of new cultivars by breeding programs requires that selected genotypes meet a wide range of characteristics, including disease resistance, productivity, and quality, among others. The selection index is a widely used biometric technique for identifying superior genotypes based on multiple traits simultaneously, but it faces reliability challenges when many attributes are involved, due to correlations among traits and the difficulty of estimating index weights. To address this issue, factor analysis (FA), a multivariate technique aimed at reducing data dimensionality and ensuring orthogonality between latent variables, has been employed. To evaluate the long-term combination of these techniques, FA scores were applied in place of phenotypic variables in the index, and this approach was compared to the exclusive use of the index. Various scenarios were simulated, with variations in the number of traits and the directions and magnitudes of their correlations, to assess the impact on selection. The efficiency of the combination was evaluated in terms of genetic gain, genetic diversity, and inbreeding, and these results were compared to the exclusive use of the selection index. Results indicate that, in the long term, the combination of techniques achieved similar genetic gains to the traditional index but with greater genetic diversity in most scenarios. The first chapter of this work is a literature review of the methods used, while the second chapter describes the execution of the study.

Keywords: Simulation; Correlation; Genetic Gain; Genetic Diversity.

SUMÁRIO

RESUMO.....	6
ABSTRACT.....	7
INTRODUÇÃO GERAL	10
CAPÍTULO 1	13
REVISÃO DE LITERATURA	13
1.1. Análise Fatorial	13
1.1.1. Autovalores e Autovetores	14
1.1.2. Adequabilidade da análise fatorial	14
1.1.2.1. O critério de Kayser-Meyer-Olkin (KMO).....	14
1.1.2.2. Teste de esfericidade de Bartlett	15
1.1.3. Estimação do número de fatores (<i>m</i>)	16
1.1.3.1. Proporção da variância explicada	16
1.1.4. Modelo Fatorial Ortogonal	17
1.1.5. Método dos Componentes Principais	17
1.1.6. Rotação Varimax	18
1.1.7. Transformação nos dados	19
1.2. Índice de seleção	19
1.2.1. Definição.....	19
1.2.2. Índice clássico de Smith e Hazel	20
1.3. Parâmetros Genéticos e Simulação de Dados	21
REFERÊNCIAS	23
CAPÍTULO 2	27
AVALIAÇÃO DO IMPACTO DA SELEÇÃO BASEADA NA COMBINAÇÃO DE ANÁLISE DE FATORES E ÍNDICES DE SELEÇÃO EM UM PROGRAMA DE MELHORAMENTO DE SELEÇÃO RECORRENTE	27
RESUMO.....	27

2.1.	INTRODUÇÃO	28
2.2.	MATERIAIS E MÉTODOS.....	30
2.2.1.	Dados Simulados	30
2.2.2.	Análise Fatorial.....	32
2.2.3.	Índices de seleção.....	34
2.2.4.	Combinação da Análise Fatorial e Índices de seleção	34
2.2.5.	Medidas de comparação entre as metodologias	34
2.2.6.	Ferramentas computacionais	35
2.3.	RESULTADOS E DISCUSSÃO.....	35
2.3.1.	Ganho genético	35
2.4.2.	Diversidade genética	48
2.4.3.	Coeficiente de Endogamia	50
2.4.4.	Número de variáveis.....	51
5.	CONCLUSÕES	53
	REFERÊNCIAS.....	54

INTRODUÇÃO GERAL

O crescimento populacional, a diminuição das áreas agricultáveis e a frequência crescente de eventos climáticos, antes considerados isolados, ameaçam a segurança alimentar global (Saath & Fachinello, 2018). Diante desse cenário, a seleção de cultivares geneticamente superiores capazes de superar esses desafios é essencial para garantir essa segurança. No entanto, para o lançamento de cultivares que atendam a essas exigências, além das demandas do mercado consumidor, é fundamental considerar uma gama de características em um programa de melhoramento, incluindo resistência a doenças, produtividade, qualidade e adaptação. Para alcançar a identificação de indivíduos ou genótipos que apresentam uma combinação de atributos ótimos para os vários caracteres alvos de um programa de melhoramento, é essencial o uso de técnicas adequadas para seleção multicaracterística que permitem a avaliação simultânea de múltiplos atributos (Torres et al., 2018).

Uma técnica biométrica eficaz para seleção simultânea de um conjunto de caracteres é o índice de seleção. Este método combina linearmente diferentes atributos, otimizando a seleção com base em um objetivo específico. Essa abordagem permite considerar múltiplos caracteres ao mesmo tempo, resultando em ganhos favoráveis em todos eles, ao invés de focar apenas em um ou em poucos atributos isolados (Cruz, 2006). Os coeficientes ou pesos do índice podem ser estimados por muitos métodos, Mulamba & Mock (1978), Smith (1936), Hazel (1943), índice de Williams (1962), entre outros, ou de acordo com o conhecimento do pesquisador, como o de Pesek & Baker (1969). Conforme mencionado por Cruz et al. (2012), os resultados da seleção via índices são superiores a seleção direta em cada característica e, apesar do ganho em cada característica ser menor, essa perda é recompensada nos ganhos de todos os caracteres. No entanto, o índice enfrenta alguns desafios em sua confiabilidade quando os caracteres envolvidos são altamente correlacionados (Cruz & Carneiro, 2003) ou quando há muitas características envolvidas no índice, tornando complexo o processo de estimar o peso para cada uma delas.

Para contornar esses desafios, uma técnica multivariada para redução da dimensionalidade do conjunto de variáveis é a Análise Fatorial (AF). Através da AF, são identificadas relações já existentes entre as características e que não são

perceptíveis diretamente, resultando na formação de novas variáveis, denominadas fatores ou variáveis latentes (Corrar et al., 2011; Ferreira, 2008; Mingoti, 2007). O modelo da AF, dependendo do algoritmo utilizado para a estimação dos fatores, pode garantir que essas novas variáveis latentes sejam ortogonais, ou seja, não correlacionadas entre si. O número de fatores (m) será menor ou igual ao número de características (ou variáveis originais), reduzindo a dimensionalidade do conjunto.

No contexto de melhoramento genético, uma das técnicas de melhoramento de populações utilizadas é a seleção recorrente. Esse método consiste em um processo cíclico e contínuo de seleção, podendo ser a nível de indivíduos ou de famílias, dentro de uma população geneticamente heterogênea, seguido de um processo de recombinação ou intercruzamento desses indivíduos selecionados por meio de critérios fenotípicos ou genótipos, para serem os pais do próximo ciclo e formar uma nova população, e assim iniciar um novo ciclo de seleção (Carena et al., 2010; Resende & Alves, 2021). A seleção recorrente tem sido amplamente empregada pelos programas de melhoramento de diversas culturas, como exemplificado pelo trabalho de Wang et al. (2024), de Cintra et al. (2023) no milho fresco e em arroz, e de Ferreira et al. (2016) no melhoramento de maracujazeiro-azedo. O ciclo de seleção recorrente continua até que a população atinja um nível satisfatório de características desejadas ou até que o método de seleção não apresente mais resultados positivos. Dessa forma, um programa de seleção recorrente envolve múltiplos ciclos, e por isso, avaliar um critério de seleção em apenas um ciclo pode levar a conclusões que não se mantêm ao longo dos ciclos subsequentes. Isso porque as técnicas ou critérios de seleção podem resultar em ganhos satisfatórios inicialmente, mas ocasionar perdas ao longo do tempo.

Diante disso, o objetivo geral do trabalho foi comparar a aplicação dos escores originados da AF no índice, substituindo as variáveis originais, visando solucionar os impasses de caracteres correlacionados e reduzir a quantidade de pesos a serem estimados. Essa abordagem foi então comparada com o uso exclusivo do índice através da simulação de um programa de seleção recorrente com 30 ciclos de melhoramento, a fim de avaliar esta combinação a longo prazo. Além disso, como objetivos específicos o estudo do impacto, para cada uma das metodologias avaliadas, do número de características fenotípicas em seleção (5, 10 e 15 características simuladas), da magnitude (alta, baixa, moderada e uma mistura das magnitudes) e do sentido (positivo, negativo e uma mistura de características com

correlações positivas e negativas) da correlação entre estes atributos. A comparação das abordagens utilizadas e o estudo do impacto de cada um dos itens (número de características, magnitude e sentido da correlação) foram avaliadas através do ganho genético, da diversidade genética e da endogamia nos ciclos de melhoramento.

Essa dissertação é dividida em dois capítulos. O capítulo 1 apresenta uma revisão da Análise Fatorial, abordando sua definição, a adequabilidade do uso da técnica por diferentes critérios: Critério de Kaiser-Meyer-Olkin (KMO) e teste de esfericidade de Bartlett, os critérios para definição do número de fatores: porcentagem da variância explicada e o *Scree-plot*. Além disso, foi retratado o modelo da AF e o método de extração dos fatores denominado Método dos componentes principais. Neste capítulo, também foi abordado a definição do índice de seleção e uma revisão sobre o índice clássico de Smith (1936) e Hazel (1943). Por fim, foram apresentadas as definições de parâmetros genéticos que podem ser utilizadas como medidas de comparação entre critérios de seleção usados em programa de melhoramento de seleção recorrente, tais como ganho genético, medida de avaliação de diversidade genética como a variância genética, e o coeficiente de endogamia. O capítulo 2 descreve detalhadamente a execução prática do trabalho, a construção da simulação e das características utilizadas, a análise dos resultados obtidos e a conclusão com as implicações para o campo do melhoramento genético.

CAPÍTULO 1

REVISÃO DE LITERATURA

1.1. Análise Fatorial

A análise multivariada refere-se a um conjunto de técnicas e métodos estatísticos voltados para a análise simultânea de múltiplas variáveis em um conjunto de dados. Conforme destacado por Mingoti (2007), esse conjunto se divide em dois grupos. O primeiro grupo engloba métodos que sintetizam a estrutura de variabilidade dos dados, atuando como ferramentas para redução da dimensionalidade do conjunto de dados. O segundo abrange técnicas de inferência estatística multivariada, como a MANOVA (Análise de Variância Multivariada) que permite testar hipóteses envolvendo várias variáveis simultaneamente.

Uma das abordagens para sintetizar a estrutura de variabilidade dos dados é a análise fatorial (AF), proposta por Spearman (1961), que tem como objetivo descrever a variabilidade dos dados originais em termos de um número possivelmente menor de variáveis não observáveis e interpretáveis, denominadas fatores (Mingoti, 2007). Os fatores são formados através do agrupamento das variáveis correlacionadas, resultando na redução da dimensionalidade do conjunto de dados sem causar perda significativa de informação. Em outras palavras, o propósito é identificar estruturas já presentes, mas que não são observáveis de forma direta (Corrar et al., 2011), e por isso, os fatores também são chamados de variáveis latentes.

Adicionalmente, mediante a análise de fatores, o número de novas variáveis é potencialmente menor, mas pode ser até igual ao número de originais. De acordo com Corrar et al. (2011), tal fenômeno é justificado na situação em que as variáveis originais atuam de maneira independente, resultando em um total de fatores equivalente ao número de variáveis originais. No entanto, se as variáveis originais são correlacionadas, pode-se concluir que existe um número menor de fatores em comparação ao número inicial de variáveis.

No modelo da análise fatorial, as variáveis originais são representadas por uma combinação linear das novas variáveis (fatores comuns) e por uma única variável latente específica (Ferreira, 2008). No entanto, antes de proceder à estimação dos parâmetros do modelo e à extração dos fatores, é necessário verificar a adequabilidade da análise fatorial e definir o número de fatores. Mas, primeiramente,

definiremos os autovalores e autovetores, que são a base de toda a teoria da análise fatorial.

1.1.1. Autovalores e Autovetores

O conhecimento dos conceitos de autovalores e autovetores é de extrema relevância na análise de fatores, sendo amplamente empregados na Estatística para redução de dimensionalidade. As variáveis originais são expressas como combinações lineares de todos os fatores, que são estimados de forma a reter o máximo de informação dos dados. Os autovalores indicam a variância explicada por cada fator, enquanto os autovetores correspondem à transformação dos dados originais. O primeiro autovalor está associado ao fator mais explicativo, o segundo ao próximo fator mais relevante, e assim sucessivamente (Regazzi & Cruz, 2020; Varella 2008). De acordo com Anton e Rorres (2012), dado uma matriz quadrada A de ordem $p \times p$, um vetor x não nulo em R^n é chamado de autovetor de A se o produto Ax for um múltiplo escalar de x , ou seja, se existir um escalar λ tal que $Ax = \lambda x$. Nesse contexto, λ é chamado de autovalor de A e x um autovetor associado a λ . A equação $(A - \lambda I)x = 0$ tem solução se, e somente se,

$$\det(A - \lambda I) = 0$$

em que I é a matriz identidade e $\det(\cdot)$ denota o determinante da matriz. Ao encontrar os autovalores λ pela resolução da equação anterior, determinam-se os respectivos autovetores x associados a cada autovalor pelo sistema homogêneo $(A - \lambda I)x = 0$.

1.1.2. Adequabilidade da análise fatorial

Ao realizar uma análise de fatores, o ajuste do modelo fundamenta-se no grau de relacionamento entre as variáveis. Dessa forma, é crucial verificar a presença de correlação entre elas, a fim de avaliar a adequação da análise. Para tal propósito, são utilizados o critério de Kayser-Meyer-Olkin (KMO) (Kaiser, 1970) e o Teste de esfericidade de Bartlett (Bartlett, 1937).

1.1.2.1. O critério de Kayser-Meyer-Olkin (KMO)

O coeficiente KMO é um índice que compara as correlações simples e parciais, sendo definido como (Kaiser, 1970):

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2}$$

em que R_{ij} é a correlação amostral entre as variáveis Y_i e Y_j , e Q_{ij} é a correlação parcial entre as variáveis Y_i e Y_j , ou seja, a correlação entre as variáveis Y_i e Y_j removendo a influência das outras variáveis.

Este índice pode assumir valores entre 0 e 1. Quando as correlações parciais se aproximam de zero, o coeficiente KMO tende a se aproximar de 1 e, de acordo com Kaiser (1970) isso equivale a dizer que a matriz de correlação inversa, $R^{-1}_{p \times p}$, está próxima de matriz diagonal.

A classificação de adequabilidade da matriz de correlação, segundo Rencher (2002), é apresentada na Tabela 1. No entanto, essa classificação pode apresentar variações dependendo da referência utilizada. Para Hair et al. (2009) e Rencher (2002), índices entre 0,5 e 1,0 são considerados aceitáveis. Já os autores Kaiser & Rice (1974) indicam que, para a adequação de ajuste de um modelo de análise fatorial, o valor de KMO deve ser maior que 0,8.

Tabela 1. Nível de adequabilidade do valor de Kaiser-Meyer-Olkin (KMO) conforme Rencher (2002).

KMO	Adequabilidade
0,9 – 1,0	Excelente
0,8 – 0,9	Ótima
0,7 – 0,8	Boa
0,6 – 0,7	Regular
0,5 – 0,6	Ruim
Abaixo de 0,5	Inadequada

1.1.2.2. Teste de esfericidade de Bartlett

O teste de esfericidade de Bartlett é útil para verificar a presença de influência entre as variáveis, ou seja, se a matriz de correlações é igual a uma matriz identidade, indicando que as variáveis são estatisticamente não correlacionadas. As hipóteses estatísticas a serem testadas são:

$$H_0: \boldsymbol{\rho}_{p \times p} = \boldsymbol{I}_{p \times p} \text{ contra } H_a: \boldsymbol{\rho}_{p \times p} \neq \boldsymbol{I}_{p \times p},$$

em que $\boldsymbol{\rho}_{p \times p}$ é a matriz de correlação das p variáveis, e $\boldsymbol{I}_{p \times p}$ é a matriz identidade de dimensão $p \times p$.

A estatística do teste é dada por:

$$T = - \left[n - \left(\frac{2p + 11}{6} \right) \right] \left[\sum_{i=1}^p \ln (\hat{\lambda}_i) \right]$$

em que $\hat{\lambda}_i$ são os autovalores estimados da matriz de correlações amostrais, n é o tamanho da amostra, e p é o número de variáveis originais. Sob a suposição da hipótese nula e para um tamanho de amostra suficientemente grande, a estatística T segue uma distribuição de qui-quadrado com $\frac{1}{2}p(p - 1)$ graus de liberdade. Para que o modelo possa ser adequadamente ajustado por meio da análise de fatores, é necessário que o teste de Bartlett seja estatisticamente significativo, ou seja, ocorra a rejeição da hipótese nula, concluindo-se assim que as características são correlacionadas (Hair et al., 2009; Maroco, 2014).

1.1.3. Estimação do número de fatores (m)

A definição do número de fatores (m) desempenha um papel crucial no sucesso da análise fatorial. Dado que o objetivo da AF é reduzir a dimensionalidade do conjunto de dados, é ideal que o valor de m seja inferior ao número de variáveis em consideração. Entretanto, isso não abrange a totalidade da variância dos dados, mas sim uma parcela que será elucidada pelos fatores. A obtenção do número de fatores pode ser realizada de diversas maneiras, incluindo procedimentos descritivos, como a proporção da variância explicada e o critério do gráfico de declive (*scree-plot*), que se baseiam na obtenção dos autovalores da matriz de correlação.

1.1.3.1. Proporção da variância explicada

Após obter os autovalores λ_i da matriz de correlação a proporção explicada por cada autovalor é dada pela razão entre cada autovalor e a soma de todos os autovalores da seguinte maneira $\frac{\lambda_j}{\sum_{j=1}^p \lambda_j}$, em que $i = 1, \dots, p$. Assim, o objetivo é selecionar o menor número de autovalores, e consequentemente, de fatores, que explique uma porcentagem satisfatória da variância dos dados. De acordo com Ferreira (2008), é recomendado que se preserve pelo menos 70% da variância total dos dados. Sendo assim, o valor de m corresponderá a quantidade de autovalores mantidos até atingir pelo menos 70% de explicação da variância do total presente nos dados, ou seja, $1 \leq m \leq p$.

Após a definição do número de fatores (m) e a confirmação da viabilidade da análise fatorial, procede-se à definição do modelo. Além dos critérios apresentados, a escolha do número de fatores, segundo Mingoti (2007), deve levar em consideração a interpretabilidade dos fatores e o princípio da parcimônia.

1.1.4. Modelo Fatorial Ortogonal

Considerando o vetor aleatório das variáveis p -dimensional $\mathbf{Y} = [Y_1, Y_2, \dots, Y_p]^T$ com média $\boldsymbol{\mu}$ ($p \times 1$) e covariância $\boldsymbol{\Sigma}$ ($p \times p$), o modelo fatorial é definido por Ferreira (2008):

$$\mathbf{Y} - \boldsymbol{\mu} = \boldsymbol{\Gamma}\mathbf{F} + \boldsymbol{\epsilon}$$

em que $\boldsymbol{\Gamma} = [\gamma_{ij}]$ é a matriz ($p \times m$) de cargas fatoriais, $\mathbf{F}$ é o vetor aleatório ($m \times 1$) de fatores comuns, e $\boldsymbol{\epsilon}$ é o vetor ($p \times 1$) de erros aleatórios ou fatores específicos.

Para realizar a estimação dos parâmetros do modelo, algumas pressuposições devem ser respeitadas, como $E(\mathbf{Y}) = \boldsymbol{\mu}$, $E(\mathbf{F}) = E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Cov(\mathbf{F}) = \mathbf{I}_m$, $Cov(\mathbf{Y}) = \boldsymbol{\Sigma}$, $Cov(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$ ($p \times p$) e $Cov(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}$. A matriz $\boldsymbol{\Psi}$ é a matriz das unicidades, com a variância específica de cada variável disposta em sua diagonal, dada por:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \psi_p \end{bmatrix}$$

com $\psi_i > 0$, para todo $i = 1, 2, \dots, p$.

De acordo com as pressuposições listadas acima, é possível verificar que $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi}$. No próximo tópico, serão apresentados um dos métodos existentes para a estimação das matrizes de cargas fatoriais ($\hat{\boldsymbol{\Gamma}}$) e das unicidades ($\hat{\boldsymbol{\Psi}}$).

1.1.5. Método dos Componentes Principais

Esta metodologia consiste na decomposição espectral da matriz de covariâncias ($\boldsymbol{\Sigma}$) ou da matriz de correlação ($\mathbf{R}$), sendo a última expressa da seguinte forma (Hongyu et al., 2016):

$$\mathbf{R} = \mathbf{P}_m \boldsymbol{\Lambda}_m \mathbf{P}_m^T$$

em que $\mathbf{P}_m$ é uma matriz ($p \times m$) formada pelos m primeiros autovetores das colunas da matriz $\mathbf{R}$ e $\boldsymbol{\Lambda}_m$ é uma matriz diagonal dos autovalores da matriz $\mathbf{R}$ associados a estes m autovetores. Desenvolvendo e manipulando as matrizes, tem-se:

$$\mathbf{R} \cong \mathbf{P}_m \boldsymbol{\Lambda}_m^{\frac{1}{2}} \boldsymbol{\Lambda}_m^{\frac{1}{2}} \mathbf{P}_m^T = \boldsymbol{\Gamma}_{p \times m} \boldsymbol{\Gamma}_{p \times m}^T.$$

E posteriormente, as unicidades ($\Psi_{p \times p}$) são obtidas por diferença. As matrizes $\hat{\Gamma}$ e $\hat{\Psi}$ serão determinadas por:

$$\hat{\Gamma} = \prod_{i=1}^m \sqrt{\hat{\lambda}_i} \hat{e}_i \quad \text{e} \quad \hat{\Psi} = \text{diag}(\mathbf{R}_{p \times p} - \hat{\Gamma} \hat{\Gamma}^T).$$

Observe que, ao empregar o método dos Componentes Principais para extrair os fatores, optou-se por considerar somente os m primeiros autovetores e autovalores da matriz $\mathbf{R}$. Além disso, no estudo descrito no Capítulo 2, utilizou-se a matriz de correlação dos dados padronizados para a decomposição espectral.

Após a estimação das matrizes $\hat{\Gamma}$ e $\hat{\Psi}$, é possível calcular os valores que serão atribuídos aos fatores $\hat{\mathbf{F}}$, denominados de escores, que são dados por (Ferreira, 2008):

$$\hat{\mathbf{F}} = \hat{\Gamma}^T (\hat{\Gamma} \hat{\Gamma}^T + \hat{\Psi})^{-1} (\mathbf{Y}_j - \bar{\mathbf{Y}})$$

em que $\mathbf{Y}_j$ é o vetor de valores observados pelo conjunto de variáveis do j -ésimo indivíduo e $\bar{\mathbf{Y}}$ é o vetor de médias referente as variáveis avaliadas.

A distribuição das variáveis originais nos fatores é determinada pelas cargas fatoriais l_{ij} (*loadings*), que representa a associação da i -ésima variável com o j -ésimo fator. O fator ao qual cada variável é alocada é determinado pelo valor absoluto mais alto das cargas fatoriais.

1.1.6. Rotação Varimax

A rotação *varimax* é um processo de rotação ortogonal aplicada aos eixos dos fatores, que facilita a identificação das variáveis associadas a cada fator ao maximizar a variância explicada pelos fatores. Isso proporciona uma melhor interpretação mais clara dos fatores e auxilia na identificação de padrões mais evidentes nos dados. A proporção de variância de cada variável que é explicada pelo fator, denominada como comunalidade, é calculada pela seguinte expressão (Ferreira, 2008):

$$c_i^2 = \sum_{j=1}^m l_{ij}^2$$

e a parte da variância de cada variável que não é elucidada pelos fatores é atribuída ao erro aleatório ou fatores específicos de cada variável, denominado também como unicidade. Alguns critérios para eliminação de variáveis com base no valor da comunalidade foram apresentados na literatura. Cruz (2006) sugere excluir uma variável da análise quando a comunalidade for inferior a 0,64, equivalentemente a uma

correlação de 0,80. Em contraste, Figueiredo e Silva (2010) defendem que variáveis com comunalidade a partir de 0,50 devem ser mantidas na análise.

1.1.7. Transformação dos dados

Ao utilizar a matriz de covariâncias Σ na estimação de Γ e Ψ , as variâncias das variáveis são influenciadas por suas escalas. Por outro lado, ao utilizar a matriz de correlações R admite-se que as variáveis têm a mesma variância (Campana et al., 2010). Essas duas abordagens têm impacto direto nos resultados da análise de fatores. Conforme proposto por Campana et al. (2010), considera-se uma transformação nas variáveis originais baseada no coeficiente de variação (CV) da seguinte forma:

$$Z_{ij}^* = Z_{ij} \times CV_j = \frac{y_{ij} - \mu_j}{\sigma_j} \times \frac{\sigma_j}{\mu_j} = \frac{y_{ij} - \mu_j}{\mu_j}$$

em que $i = 1, \dots, n$ observações, $j = 1, \dots, p$ variáveis, Z_{ij}^* é o valor da i -ésima observação da variável Y_j padronizada, y_{ij} é a i -ésima observação da j -ésima variável, μ_j é a média da j -ésima variável e σ_j é o desvio padrão da j -ésima variável.

O coeficiente de variação foi utilizado porque esta medida é a mais adequada para descrever a variabilidade relativa das variáveis, uma vez que não possui unidade de medida. Pela equação anterior, percebe-se que Z_{ij}^* segue distribuição normal multivariada, com $E(Z_{ij}^*) = 0$ e $V(Z_{ij}^*) = CV_i^2$. Portanto, as variáveis transformadas são ajustadas para uma mesma escala de medida, com médias iguais a zero e variâncias diferentes de um, mas proporcionais a seus respectivos coeficientes de variação.

1.2. Índice de seleção

1.2.1. Definição

O principal objetivo dos programas de melhoramento é selecionar indivíduos geneticamente superiores que atendam às necessidades comerciais, tanto dos consumidores quanto dos produtores. Para alcançar esse objetivo, o material genético selecionado deve satisfazer simultaneamente diversos atributos desejáveis da variedade. Neste contexto, a aplicação da técnica biométrica de índice de seleção se mostra eficiente, pois possibilita, por meio de uma combinação ótima de variáveis a seleção simultânea de vários caracteres (Cruz et al., 2012). O índice de seleção I é

definido como uma combinação linear das características fenotípicas que são alvos do programa de melhoramento, e pode ser estabelecido da seguinte forma:

$$I = b_1y_1 + b_2y_2 + \dots + b_py_p$$

em que os b_i 's são os coeficientes e os y_i 's são as variáveis para as quais se deseja selecionar. O peso atribuído a cada característica pode ser determinado de diversas maneiras, refletindo tanto a subjetividade do melhorista, que pode favorecer características específicas que deseja priorizar nas próximas gerações, quanto a utilização de métodos de otimização ou parâmetros genéticos. Além disso, os pesos podem ter uma conotação econômica, sendo então denominados pesos econômicos.

1.2.2. Índice clássico de Smith e Hazel

A construção dos índices de seleção envolve a estimação dos pesos ou coeficientes de ponderação associados ao índice de seleção. O índice clássico proposto por Hazel (1943) e Smith (1936) sugere que os pesos possam ser estimados maximizando a correlação entre o índice e o agregado genotípico.

Considere o índice (I), o agregado genotípico (H), que é determinado através de uma combinação linear dos valores genéticos ponderados por seus respectivos pesos, e a correlação entre eles com n caracteres avaliados, temos (Cruz et al., 2012):

$$I = b_1y_1 + b_2y_2 + b_3y_3 + \dots + b_py_p = \mathbf{b}'\mathbf{y}$$

$$H = a_1g_1 + a_2g_2 + a_3g_3 + \dots + a_pg_p = \mathbf{a}'\mathbf{g}$$

$$\rho_{IH} = \frac{\mathbf{b}'\mathbf{G}\mathbf{a}}{\sqrt{(\mathbf{b}'\mathbf{P}\mathbf{b})(\mathbf{a}'\mathbf{G}\mathbf{a})}}$$

em que $\mathbf{b}' = [b_1, b_2, \dots, b_p]$ é o vetor dos coeficientes de ponderação do índice de dimensão $1 \times p$, $\mathbf{a}' = [a_1, a_2, \dots, a_p]$ é o vetor dos pesos do agregado genotípico previamente estabelecidos de dimensão $1 \times p$, $\mathbf{G}$ é a matriz de covariâncias genéticas ($p \times p$) e $\mathbf{P}$ é a matriz de covariâncias fenotípicas ($p \times p$) entre os caracteres.

Para obter os pesos, a maximização pode ser realizada na correlação ou mais facilmente por meio da função $\ln(\rho)$, uma vez que a função logarítmica é monótona crescente. Dessa forma, tem-se:

$$\hat{\mathbf{b}} = \mathbf{P}^{-1}\mathbf{G}\mathbf{a}$$

Entretanto, a técnica de seleção de índice enfrenta um grande desafio quando há um grande número de variáveis a serem consideradas no índice. A existência de correlação entre os caracteres, denominada multicolinearidade, pode acarretar resultados equivocados em muitos procedimentos biométricos, incluindo a técnica de índice de seleção. Isso ocorre porque a matriz ***P*** precisa ser invertível e na presença de correlação entre as variáveis, a inversão pode não ser possível. A confiabilidade na estimativa dos coeficientes de ponderação é limitada, tornando-os inadequados para uso na seleção em um programa de melhoramento genético (Cruz & Carneiro, 2003).

1.3. Parâmetros Genéticos e Simulação de Dados

Os parâmetros genéticos, como por exemplo o ganho genético, a diversidade genética e a endogamia, são quantidades estimadas a partir de dados de uma população, permitindo sua caracterização. O melhorista monitora constantemente esses parâmetros em suas populações no campo para orientar e acompanhar as decisões tomadas. Do ponto de vista estatístico, o pesquisador pode utilizá-los para avaliar as técnicas estatísticas e genéticas empregadas na seleção. Além disso, outra ferramenta que pode auxiliar o melhorista a orientar suas decisões antes da implementação no campo, evitando gastos desnecessários de trabalho e recursos financeiros, é a avaliação das técnicas em dados simulados que reproduzam seu programa de melhoramento. Somente em cenários simulados é possível conhecer as estimativas ótimas dos parâmetros genéticos da população e avaliar o quão próximas as estimativas dos parâmetros associados a cada metodologia estarão desses valores, e o quão acurada está sendo o processo de seleção. Atualmente, a simulação de dados tem sido amplamente aplicada na agricultura, destacando-se os trabalhos de Batista et al. (2021), Lima et al. (2022), Azevedo et al. (2024), Jannink et al. (2023) e Fritsche-Neto et al. (2024). Softwares específicos para simulação também se tornaram disponíveis ao longo dos anos, como o AlphaSimR (Chris Gaynor et al., 2021) e o FieldSimR (Werner et al., 2024).

Os parâmetros genéticos geralmente monitorados pelo melhorista e que são utilizados na maioria dos estudos de simulação, como nos trabalhos citados anteriormente, incluem:

i) Ganho genético: O ganho genético mede a superioridade genética dos descendentes em relação à geração anterior (Carneiro Junior, 2009).

ii) Diversidade Genética: A diversidade genética avalia a quantidade de variação presente na espécie de interesse e pode ser medida por meio da variância genética dos pais a cada ciclo (Cruz et al., 2011).

iii) Endogamia: A endogamia refere-se ao acasalamento de indivíduos aparentados, que são geneticamente semelhantes e o resultado desse acasalamento é o aumento da homozigose (Lôbo, 2008). Devido ao aumento da homozigose, a endogamia aumenta a semelhança entre os indivíduos e, conseqüentemente, reduz a diversidade genética entre eles. A endogamia pode ser medida por meio da média da diagonal da matriz de parentesco entre os pais, subtraindo 1 (Resende et al., 2014).

REFERÊNCIAS

- Anton, H., & Rorres, C. (2012). **Álgebra linear com aplicações** (10.ed.). Porto Alegre: Bookman. p. 768.
- Azevedo, C. F., Ferrão, L. F. V., Benevenuto, J., de Resende, M. D. V., Nascimento, M., Nascimento, A. C. C., & Munoz, P. R. (2024). Using visual scores for genomic prediction of complex traits in breeding programs. **Theoretical and Applied Genetics**, 137(1). <https://doi.org/10.1007/s00122-023-04512-w>
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901), 268–282.
- Batista, L. G., Gaynor, R. C., Margarido, G. R. A., Byrne, T., Amer, P., Gorjanc, G., & Hickey, J. M. (2021). Long-term comparison between index selection and optimal independent culling in plant breeding programs with genomic prediction. **PLoS ONE**, 16(5 May 2021). <https://doi.org/10.1371/journal.pone.0235554>
- Campana, A. C. M., Ribeiro Júnior, J. I., & Nascimento, M. (2010). Uma proposta de transformação de dados para a análise de componentes principais. **Revista Brasileira de Biometria**, 28(2), 103–115.
- Carena, M. J., Hallauer, A. R., Miranda Filho, J. B., Hallauer, A. R., Carena, M. J., & Filho, J. B. M. (2010). Selection: Theory. **Quantitative Genetics in Maize Breeding**, 223–290.
- Carneiro Junior, J. M. (2009). **Melhoramento genético animal**. In: GONCALVES, R. C.; DE OLIVEIRA, L. C. *Embrapa Acre: ciência e tecnologia para o desenvolvimento sustentável do Sudoeste da Amazônia*. Rio Branco, AC: Embrapa Acre, 2009. cap. 11, p. 197-208.
- Chris Gaynor, R., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: An R package for breeding program simulations. **G3: Genes, Genomes, Genetics**, 11(2). <https://doi.org/10.1093/G3JOURNAL/JKAA017>
- Cintra, P. H. N., Resende, C. L. P., Damaso, L. F., Carvalho, D. D. C., Silva, F. de C., & Rodrigues, F. (2023). Five cycles of intrapopulation recurrent selection in half-sib progenies of fresh corn. **Revista Caatinga**, 36(3), 723–730. <https://doi.org/10.1590/1983-21252023v36n324rc>
- Corrar, L., Paulo, E., Dias Filho, J. M., & Rodrigues, A. (2011). **Análise multivariada para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas.
- Cruz, C. D., & Carneiro, P. C. S. (2003). **Modelos biométricos aplicados ao melhoramento de plantas**. Viçosa: UFV.
- Cruz, C. D. (2006). **Modelos biométricos aplicados ao melhoramento genético**. UFV.

Cruz, C. D. (2006). **Programa GENES: análise multivariada e simulação**. Viçosa: Editora UFV.

Cruz, C. D., Ferreira, F. M., & Pessoni, L. A. (2011). **Biometria aplicada ao estudo da diversidade genética**. Suprema.

Cruz, C. D., Regazzi, A. J., & Carneiro, P. C. S. (2012). **Modelos biométricos aplicados ao melhoramento**. Viçosa: UFV.

Ferreira, D. F. (2008). **Estatística multivariada**. Editora Ufla Lavras.

Ferreira, R. T., Viana, A. P., Silva, F. H. de L. E., Santos, E. A., & Santos, J. O. (2016). Intrapopulation recurrent selection in sour passion fruit by mixed models. **Revista Brasileira de Fruticultura**, 38(1), 158–166. <https://doi.org/10.1590/0100-2945-260/14>

Figueiredo Filho, D. B & Júnior, J. A. S. **Visão além do alcance: uma introdução à análise fatorial**. Opinião Pública, Campinas, v. 16, n. 1, p. 160-185, 2010.

Fritsche-Neto, R., Ali, J., De Asis, E. J., Allahgholipour, M., & Labroo, M. R. (2024). Improving hybrid rice breeding programs via stochastic simulations: number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. **Theoretical and Applied Genetics**, 137(1). <https://doi.org/10.1007/s00122-023-04508-6>

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). **Análise multivariada de dados**. Bookman editora.

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. **Genetics**, 28(6), 476–490.

Hongyu, K., Sandanielo, V. L. M., & de Oliveira Junior, G. J. (2016). Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, 5(1), 83-90.

Jannink, J. L., Astudillo, R., & Frazier, P. (2023). Insight into a two-part plant breeding scheme through Bayesian optimization of budget allocations. **Crop Science**. <https://doi.org/10.1002/csc2.21124>

Kaiser, H. F. (1970). A second generation little jiffy. **Psychometrika**, 35, 401-415. <https://doi.org/10.1007/BF02291817>

Kaiser, H. F. (1974). An index of factorial simplicity. **Psychometrika**, 39(1), 31-36. <https://doi.org/10.1007/BF02291575>

Lima, L. P., Azevedo, C. F., de Resende, M. D. V., Nascimento, M., & Fonseca E Silva, F. (2022). Evaluation of bayesian methods of genomic association via chromosomic regions using simulated data. **Scientia Agricola**, 79(3). <https://doi.org/10.1590/1678-992X-2020-0202>

Lôbo, R. (2008). **Melhoramento genético de ovinos e caprinos**. <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/534093/1/AAC-Melhoramento-genetico-de-ovinos.pdf>

Maroco, J. (2014). Análise estatística com o SPSS. **Statistics**, 6, 370.

Mingoti, S. A. (2007). **Análise de dados através de métodos estatística multivariada: uma abordagem aplicada** (p. 295).

Mulamba, N. N., & Mock, J. J. (1978). *Improvement of yield potential of the Eto Blanco maize (Zea mays L.) population by breeding for plant traits*. **Egypt J Genet Cytol** 7:40–51

Pesek, J., & Baker, R. J. (1969). Desired improvement in relation to selected indices. **Can. J. Plant. Sci.**, 49, 803–804.

Regazzi, A. J., Cruz, C. D. (2020). **Análise Multivariada Aplicada**. Viçosa, MG, p. 1-401.

Rencher, A. C. (2002). **Methods of Multivariate Analysis**, 2nd edn., A John Wiley & Sons. *New York*.

Resende, M. de, & Alves, R. (2021). **Genética: estratégias de melhoramento e métodos de seleção**. <https://www.alice.cnptia.embrapa.br/alice/handle/doc/1131863>

Resende, M. de, Silva, F. F., & Azevedo, C. F. (2014). **Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa: UFV.

Saath, K. C. de O., & Fachinello, A. L. (2018). Crescimento da Demanda Mundial de Alimentos e Restrições do Fator Terra no Brasil. **Revista de Economia e Sociologia Rural**, 56(2), 195–212. <https://doi.org/10.1590/1234-56781806-94790560201>

Smith, H. F. (1936). A discriminant function for plant selection. **Annals of Eugenics**, 7(3), 240–250. <https://doi.org/10.1111/j.1469-1809.1936.tb02143.x>

Spearman, C. (1961). “General Intelligence” Objectively Determined and Measured. **American Journal of Psychology**, 100(3), 697.

Team, R. C. (2021). R: A language and environment for statistical computing. URL. **R Foundation for Statistical Computing: Vienna, Austria**.

Teixeira, F. R. F., Nascimento, M., Nascimento, A. C. C., e Silva, F. F., Cruz, C. D., Azevedo, C. F., Paixão, D. M., Barroso, L. M. A., Verardo, L. L., de Resende, M. D. V., Guimarães, S. E. F., & Lopes, P. S. (2016). Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research**, 15(2). <https://doi.org/10.4238/gmr.15028231>

Torres, L. G., Rodrigues, M. C., Lima, N. L., Horta Trindade, T. F., Fonseca e Silva, F., Azevedo, C. F., & DeLima, R. O. (2018). Multi-trait multi-environment Bayesian model reveals g x e interaction for nitrogen use efficiency components in tropical maize. **PLoS ONE**, 13(6). <https://doi.org/10.1371/journal.pone.0199492>

Varella, C. A. A. (2008). **Análise de componentes principais**. Departamento de Engenharia da Universidade Federal Rural do Rio de Janeiro, Seropédica – RJ.

Wang, R., Axtman, J., Leng, Y., Salsman, E., Hegstad, J., Fiedler, J. D., Xu, S., Zhong, S., Elias, E., & Li, X. (2024). Recurrent selection for Fusarium head blight resistance in a durum wheat population. **Crop Science**.

Werner, C. R., Gemenet, D. C., & Tolhurst, D. J. (2024). FieldSimR: an R package for simulating plot data in multi-environment field trials. **Frontiers in Plant Science**, 15. <https://doi.org/10.3389/fpls.2024.1330574>

Williams, J. (1962). The evaluation of a selection index. **Biometrics**, 18(3), 375–393.

CAPÍTULO 2

AVALIAÇÃO DO IMPACTO DA SELEÇÃO BASEADA NA COMBINAÇÃO DE ANÁLISE DE FATORES E ÍNDICES DE SELEÇÃO EM UM PROGRAMA DE MELHORAMENTO DE SELEÇÃO RECORRENTE

RESUMO

A seleção de um novo genótipo em programas de melhoramento exige a consideração de uma ampla gama de características de interesse, cuja combinação é fundamental para determinar o sucesso na prática agrícola e a aceitação no mercado. Nesse contexto, os índices de seleção são ferramentas eficazes para selecionar múltiplas características simultaneamente, combinando-as linearmente para obter ganhos simultâneos. No entanto, a estimação dos pesos do índice pode ser complexa quando muitos caracteres estão envolvidos, devido à multicolinearidade, o que afeta a confiabilidade do índice. A análise fatorial é uma técnica estatística multivariada que reduz a dimensionalidade do conjunto de dados, agrupando as variáveis e transformando-as em novas variáveis latentes, denominadas fatores. O modelo ortogonal da análise de fatores estima fatores ortogonais, ou seja, não correlacionados. O objetivo deste trabalho foi avaliar o uso dos escores da análise fatorial, em substituição às variáveis fenotípicas no índice de seleção, em um programa de melhoramento a longo prazo. Uma simulação foi realizada para analisar a combinação dessas técnicas na seleção dos pais ao longo de 30 ciclos de melhoramento. Os parâmetros genéticos utilizados para comparar a combinação das técnicas com o uso exclusivo do índice foram o ganho genético, a diversidade genética e a endogamia. Os resultados indicaram que, em conjuntos de características correlacionadas positivamente ou de forma mista, a combinação das técnicas obteve ganhos genéticos idênticos ao uso exclusivo do índice de seleção, além de maior diversidade genética, independentemente das correlações entre as características. A endogamia foi acentuada em todos os cenários em que as técnicas foram combinadas. De modo geral, o melhor desempenho em ganho e diversidade genética foi observado no cenário com cinco características ao empregar a combinação das técnicas.

Palavras-chave: Correlação; Simulação; Ganho Genético; Diversidade Genética; Endogamia.

2.1. INTRODUÇÃO

Os programas de melhoramento vegetal desempenham um papel crucial na seleção de genótipos ao longo das diversas etapas do desenvolvimento de cultivares. A escolha de uma variedade geneticamente superior demanda a consideração de uma ampla gama de características fenotípicas de interesse, não se restringindo a uma ou duas características isoladas, mas abrangendo um conjunto diversificado de atributos essenciais tanto para a seleção bem-sucedida quanto para aplicações comerciais específicas (Paixão et al., 2022; Torres et al., 2018). O estudo das relações entre essas características, incluindo sua natureza e magnitude, é de suma importância para os programas de melhoramento genético (Borém et al., 2021), uma vez que alterações em uma característica podem afetar outras, como observado no estudo de Cabral et al. (2016) e Entringer et al. (2014).

Uma técnica eficiente para a seleção simultânea de múltiplos caracteres é o de índice de seleção. Essa metodologia consiste em estabelecer um novo carácter que combina as múltiplas características consideradas nos indivíduos avaliados, possibilitando ganhos favoráveis e simultâneos (Cruz & Regazzi, 1994). No entanto, essa abordagem enfrenta desafios quando há muitas características devido à multicolinearidade, que afeta a confiabilidade do índice (Cruz & Carneiro, 2003). No caso do índice de Smith-Hazel (Hazel, 1943; Smith, 1936), a inversão de uma matriz de covariância fenotípica pode resultar em matrizes mal condicionadas e estimativas viesadas de ganhos genético na presença de multicolinearidade (Olivoto & Nardino, 2020). A correlação entre características pode ser causada por efeitos pleiotrópicos dos genes, que influenciam várias características fenotípicas, ou por desequilíbrios de ligação, sendo essa correlação genética transitória, especialmente nas primeiras gerações de populações obtidas pelo cruzamento de genitores geneticamente divergentes (Falconer, 1964; Vencovsky, 1978).

Islam et al. (2016) afirmam que a eficiência do índice é maior do que na seleção individual, sendo sua superioridade mais evidente à medida que o número de características sob seleção aumenta (Vachhani et al., 2016). Nesse cenário, a análise de fatores (AF) emerge como uma solução viável para simplificar o uso de índices de seleção, reduzindo a dimensionalidade do conjunto de dados e criando variáveis latentes (Ferreira, 2008). Teixeira et al. (2016) propuseram a aplicação da análise de fatores a fim de reduzir a dimensionalidade e realizar a predição genômica para grupos de características de suínos. Fialho et al. (2023) também aplicaram este método em

características fenotípicas do arroz *Oryza sativa* para reduzir o número de marcadores moleculares associados a grupos de características na predição genômica. Uma das abordagens para construir o modelo da AF presume a ortogonalidade dos fatores, resultando em variáveis não correlacionadas. A combinação das duas abordagens, AF e índices de seleção, facilitará a estimação dos coeficientes de ponderação ao eliminar a multicolinearidade (Ferreira et al., 2005). Ferreira et al. (2005) avaliaram a combinação da análise de fatores e do índice de seleção para a seleção de indivíduos em dados de café.

No entanto, a aplicação de critérios de seleção em programas de seleção recorrente avaliados em apenas um ciclo pode levar a conclusões que não se mantêm ao longo dos ciclos subsequentes. As técnicas ou critérios podem inicialmente funcionar bem, mas ao longo do tempo podem resultar em perdas. Por isso, a simulação pode orientar os melhoristas na tomada de decisões antes da implementação no campo, evitando gastos desnecessários de trabalho e recursos financeiros. Somente em cenários simulados é possível conhecer os verdadeiros parâmetros genéticos da população e avaliar quão próximas as estimativas dos parâmetros associados a cada metodologia estarão desses valores. Atualmente, a simulação de dados tem sido amplamente aplicada na agricultura, como evidenciado nos trabalhos de Azevedo et al. (2024), Batista et al. (2021), Fritsche-Neto et al. (2024), Jannink et al. (2023), Lima et al. (2022).

Diante do exposto, o presente estudo teve como objetivo geral comparar a aplicação dos escores originados da AF no índice, substituindo as variáveis originais, visando solucionar os problemas causados por características correlacionadas e reduzir a quantidade de pesos a serem estimados. Essa abordagem foi então comparada com o uso exclusivo do índice através da simulação de um programa de seleção recorrente com 30 ciclos de melhoramento, a fim de avaliar esta combinação a longo prazo. Além disso, como objetivos específicos, o estudo investigou o impacto do número de características fenotípicas em seleção (5, 10 e 15 características simuladas), da magnitude (alta, baixa, moderada e uma mistura das magnitudes) e do sentido (positivo, negativo e uma mistura de características com correlações positivas e negativas) da correlação entre estes atributos. As abordagens foram comparadas em termos de ganho genético, diversidade genética e endogamia nos ciclos de melhoramento.

2.2. MATERIAIS E MÉTODOS

2.2.1. Dados Simulados

O conjunto de dados utilizado neste estudo originaram-se da simulação de um programa de melhoramento de seleção recorrente. Essa simulação foi realizada por meio do software R (R Core Team, 2021), utilizando o pacote AlphaSimR (Faux et al., 2016). O esquema da simulação foi inspirado no trabalho de Batista et al. (2021) e está apresentado na Figura 1.

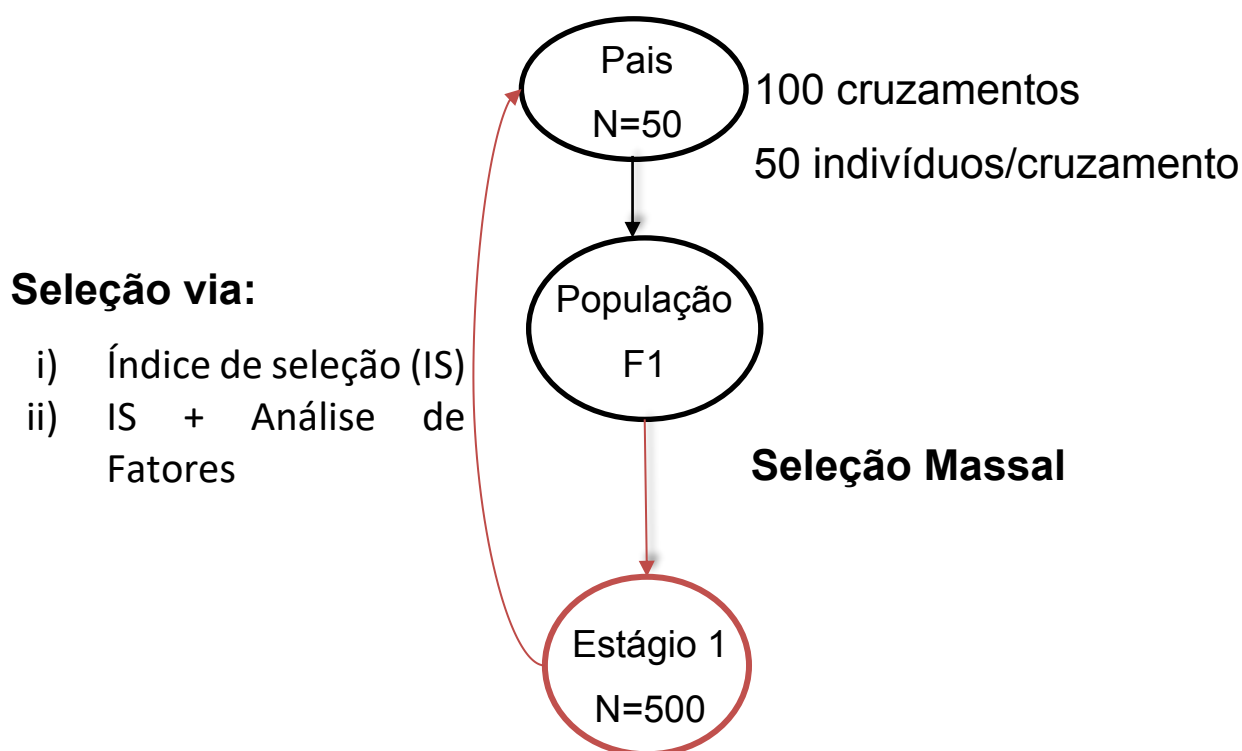


Figura 1. Esquema da simulação de um programa de melhoramento de seleção recorrente e tamanho das populações simuladas.

A simulação foi dividida em duas fases: burn-in e ciclos de melhoramento. Inicialmente, 50 genitores foram simulados, dos quais 100 cruzamentos foram selecionados aleatoriamente na primeira geração da fase de burn-in. De cada cruzamento, originaram-se 50 descendentes, totalizando 5.000 indivíduos na população F1. No estágio 1, aplicou-se a seleção massal, selecionando 500 indivíduos para fenotipagem de 5, 10 e 15 características distintas, respectivamente, proporcionando a criação de cenários que refletissem os conjuntos variados de características para a avaliação das metodologias multivariadas. Nos ciclos de burn-

in, os 50 pais selecionados, para o próximo ciclo, foram selecionados por meio do índice de seleção. No entanto, para os ciclos de melhoramento, tanto o índice de seleção quanto a combinação do índice de seleção com a análise de fatores foram utilizados para a seleção dos pais. O processo consistiu em 15 ciclos de burn-in e 30 ciclos de melhoramento, com cada cenário simulado 50 vezes.

Destaca-se que, embora as características exibam diferentes níveis de herdabilidade, variando de baixa a alta magnitude, mantivemos essa herdabilidade uniforme entre os cenários com diferentes números de características. A herdabilidade foi definida de forma a mimetizar um programa de melhoramento da seguinte maneira: i) 2 características de baixa (entre 0 e 0.30), 2 de média (entre 0.3 e 0.7) e 1 de alta (entre 0.70 e 1) herdabilidades; ii) 5 características de baixa, 4 de média e 1 de alta herdabilidades; iii) 7 características de baixa, 5 de média e 3 de alta herdabilidades para todos os cenários com 5, 10 e 15 caracteres, respectivamente. Tal abordagem garantiu uma análise consistente e comparável entre os conjuntos de características considerados. Todas as características são controladas por 120 loci de características quantitativas (*Quantitative Trait Locus* – QTL).

Além disso, em cada cenário de número de características, foram investigadas as magnitudes das correlações entre essas características. Uma variedade de cenários foi estabelecida, abrangendo correlações positivas, negativas e mistas. Dentro de cada tipo de correlação, foram examinados cenários com diferentes níveis de intensidade, que variavam desde correlações altas (-1 a -0.70 e 0.70 a 1) até baixas (-0.30 a 0 e 0 a 0.30), incluindo também correlações moderadas (-0.60 a -0.40 e 0.40 a 0.60) e mistas (incluindo todos os níveis de correlação) escolhidas aleatoriamente entre os intervalos. A criação detalhada dos cenários possibilitou uma avaliação minuciosa do impacto das correlações entre as características ao longo dos anos no processo de seleção dos pais usando as duas abordagens multivariadas. Resumidamente, tem-se três cenários de correlação (positiva, negativa e positiva/negativa). Dentro de cada um, há subcenários com diferentes intensidades alta, moderada, baixa e mista), e ainda subcenários com 5, 10 e 15 características.

Nos ciclos de melhoramento, a seleção dos pais foi conduzida por meio de duas abordagens distintas: i) o índice de seleção de Smith e Hazel, no qual o vetor de pesos foram assumidos igualmente para todos as características; e ii) o índice de seleção de Smith e Hazel baseado nos escores provenientes da análise de fatores. Após a

seleção utilizando essas diferentes metodologias, os indivíduos selecionados serviram como genitores para o próximo ciclo reprodutivo.

Ao término de cada ciclo, foi realizada uma avaliação e comparação entre cada uma das metodologias em relação às demais. As medidas avaliativas foram a média dos valores obtidos nas replicatas.

Tabela 2. Descrição dos cenários construídos via simulação.

Cenários de simulação construídos	
Sinal da correlação	Magnitude da correlação
Correlação Positiva	Alta
	Baixa
	Mista
	Moderada
Correlação Positiva Negativa	Alta
	Baixa
	Mista
	Moderada
Correlação Negativa	Alta
	Baixa
	Mista
	Moderada

2.2.2. Análise Fatorial

As variáveis no modelo da análise fatorial são representadas por uma combinação linear das variáveis latentes, os fatores, e pelo erro aleatório ou fator específico para cada variável (Ferreira, 2008):

$$Y - \mu = \Gamma F + \epsilon$$

em que Y é o vetor aleatório p -dimensional, μ é o vetor de médias ($p \times 1$), Γ é a matriz ($p \times m$) de cargas fatoriais (l_i , $i = 1, \dots, p$ variáveis), F é o vetor aleatório ($m \times 1$) de fatores comuns latentes e ϵ é um vetor ($p \times 1$) de erros aleatórios ou fatores específicos.

Considerando que um dos objetivos específicos do estudo foi avaliar o impacto das correlações na análise, tanto em termos de direção quanto de magnitude, optou-

se por não verificar a adequabilidade, conforme orientado por Kaiser (1970). O número de fatores foi estabelecido de acordo com o critério da porcentagem da variância explicada (Ferreira, 2008), definida como $\frac{\lambda_j}{\sum_{j=1}^p \lambda_j}$, sendo o número de fatores suficiente para explicar pelo menos 75% da variabilidade dos dados.

Para estimar o vetor de fatores comuns (F), é necessário primeiro estimar as matrizes de cargas fatoriais (Γ) e das unicidades (Ψ). A estimação dessas matrizes é realizada pela decomposição espectral da matriz de covariâncias (Σ) ou da matriz de correlação (R). No entanto, ao utilizar Σ , a variância das variáveis é influenciada por suas escalas, e ao utilizar R , assume-se que as variáveis têm a mesma variação. Portanto, admite-se uma transformação nas variáveis originais por meio do coeficiente de variação (CV) da seguinte forma (Campana et al., 2010):

$$Z_{ij}^* = Z_{ij} \times CV_j = \frac{y_{ij} - \mu_j}{\sigma_j} \times \frac{\sigma_j}{\mu_j} = \frac{y_{ij} - \mu_j}{\mu_j}$$

em que $i = 1, \dots, n$ observações, $j = 1, \dots, p$ variáveis, Z_{ij}^* é o valor da i -ésima observação da variável Y_j padronizada, y_{ij} é a i -ésima observação da Y_j variável, μ_j é a média da Y_j variável e σ_j é o desvio padrão da Y_j variável. Conforme proposto por Campana et al. (2010), o coeficiente de variação é a medida de dispersão dos dados mais adequada para descrever a variação dos dados por ser adimensional.

A estimação das matrizes de cargas fatoriais e unicidades dos dados transformados é dada por:

$$\hat{\Gamma} = \prod_{i=1}^m \sqrt{\hat{\lambda}_i} \hat{e}_i \quad \text{e} \quad \hat{\Psi} = \text{diag}(R_{p \times p} - \hat{\Gamma} \hat{\Gamma}^T)$$

onde $\hat{\lambda}_i$ é o i -ésimo autovalor e $\hat{e}_i$ é o i -ésimo autovetor da matriz dos dados transformados. Posteriormente, os escores dos fatores são obtidos por (Ferreira, 2008):

$$\hat{F} = \hat{\Gamma}^T (\hat{\Gamma} \hat{\Gamma}^T + \hat{\Psi})^{-1} (Y_j - \bar{Y})$$

em que Y_j é vetor das características fenotípicas transformadas, $\bar{Y}$ vetor de médias das variáveis transformadas, $\hat{\Gamma}$ matriz de cargas fatoriais e $\hat{\Psi}$ matriz de unicidades estimadas anteriormente.

$$c_i^2 = \sum_{j=1}^m l_{ij}^2$$

2.2.3. Índices de seleção

O índice I é representado como uma combinação linear das características fenotípicas da seguinte forma:

$$I = b_1y_1 + b_2y_2 + \dots + b_py_p$$

em que os b_i e y_i com $i = 1, \dots, p$ são os coeficientes e as variáveis para as quais se deseja selecionar, respectivamente.

O índice de Smith (1936) e Hazel (1943) estima os coeficientes maximizando a correlação do índice com uma combinação linear dos valores genéticos ponderados por seus respectivos coeficientes, obtendo (Cruz et al., 2012):

$$\hat{\mathbf{b}} = \mathbf{P}^{-1}\mathbf{G}\mathbf{a}$$

em que $\mathbf{a}$ é o vetor dos pesos do agregado genotípico previamente estabelecidos de dimensão $p \times 1$, $\mathbf{G}$ é a matriz de covariâncias genéticas ($p \times p$) e $\mathbf{P}$ é a matriz de covariâncias fenotípicas ($p \times p$) entre os caracteres.

2.2.4. Combinação da Análise Fatorial e Índices de seleção

Ao realizar a análise de fatores nas características fenotípicas, espera-se que o número de variáveis latentes formadas seja menor do que anteriormente, incorporando-as no índice de seleção. O índice de seleção foi definido com base nos escores dos fatores associados a grupos de características e é dado por:

$$I = b_1F_1 + b_2F_2 + b_3F_3 + \dots + b_pF_p$$

em que os b_i 's são os pesos a serem estimados e os F_i 's são os escores associados a cada um dos fatores. Os pesos foram definidos por meio das expressões como a Smith (1936), Hazel (1943) e Pesek e Baker (1969), mas agora sem a presença de multiconearidade devido a ortogonalidade dos fatores.

2.2.5. Medidas de comparação entre as metodologias

Após a seleção dos pais pelas metodologias - a combinação da análise de fatores e índices de seleção e a seleção apenas pelo índice - a comparação entre elas foi realizada da seguinte forma: i) Ganho genético: Avaliado pela média do valor genético paramétrica dos pais a cada ciclo; ii) Diversidade Genética: Medida pela variância genética paramétrica dos pais a cada ciclo; iii) Endogamia: Calculada pela

média da diagonal da matriz de parentesco entre os pais, subtraindo 1 (Resende et al., 2014).

2.2.6. Ferramentas computacionais

A análise fatorial foi conduzida com a função *principal* do pacote *psych* (Revelle, 2019), enquanto a elaboração do índice de Smith e Hazel foi realizada através da função *smithHazel* do pacote *AlphaSimR* (Faux et al., 2016), ambas implementadas no software *R* (R Core Team, 2021).

2.3. RESULTADOS E DISCUSSÃO

2.3.1. Ganho genético

O ganho genético foi mais evidente em cenários em que as características apresentavam correlações positivas (Figuras 2, 3, 4 e 5). Esses resultados estão alinhados com o trabalho de Bajgain et al. (2019) e Paixão et al. (2022), que destacam a possibilidade de melhorar simultaneamente características correlacionadas positivamente desde que se pretende aumentar ou diminuir todas as características de forma conjunta. No cenário de correlações positivas e negativas entre si (cenário misto - Figuras 6, 7, 8 e 9), o ganho foi inferior ao de correlações exclusivamente positivas, porém maior que em características correlacionadas negativamente (Figuras 10, 11, 12 e 13). Conforme ilustrado por Yao et al. (2018), a correlação negativa entre rendimento e qualidade limitou o ganho genético obtido na seleção genômica em trigo.

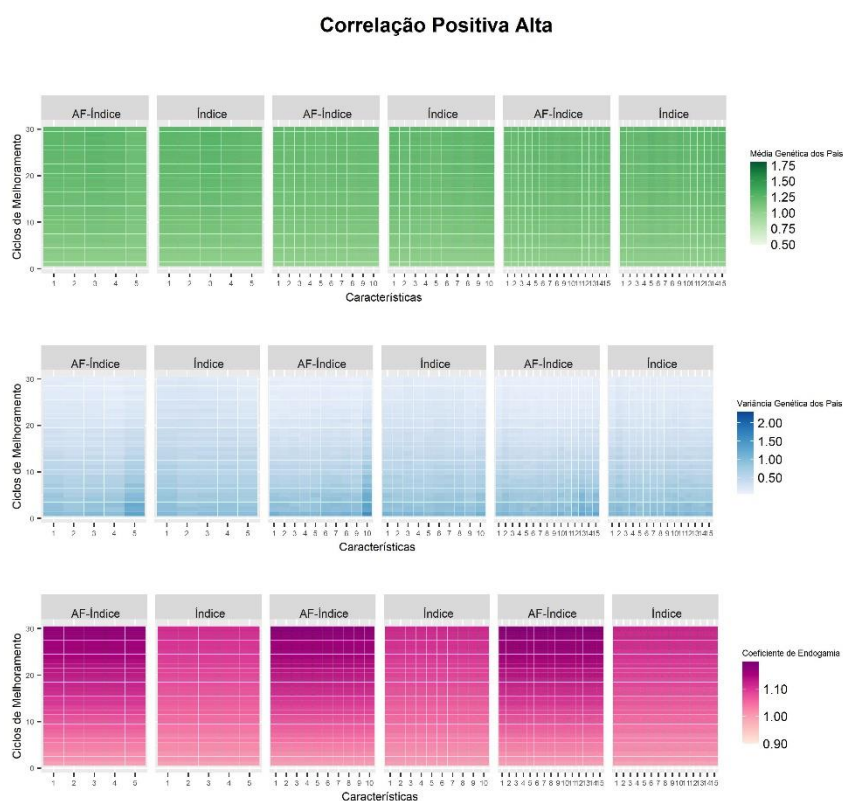


Figura 2. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5$, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas são altamente positivamente correlacionadas (correlação variando entre 0.70 e 1).

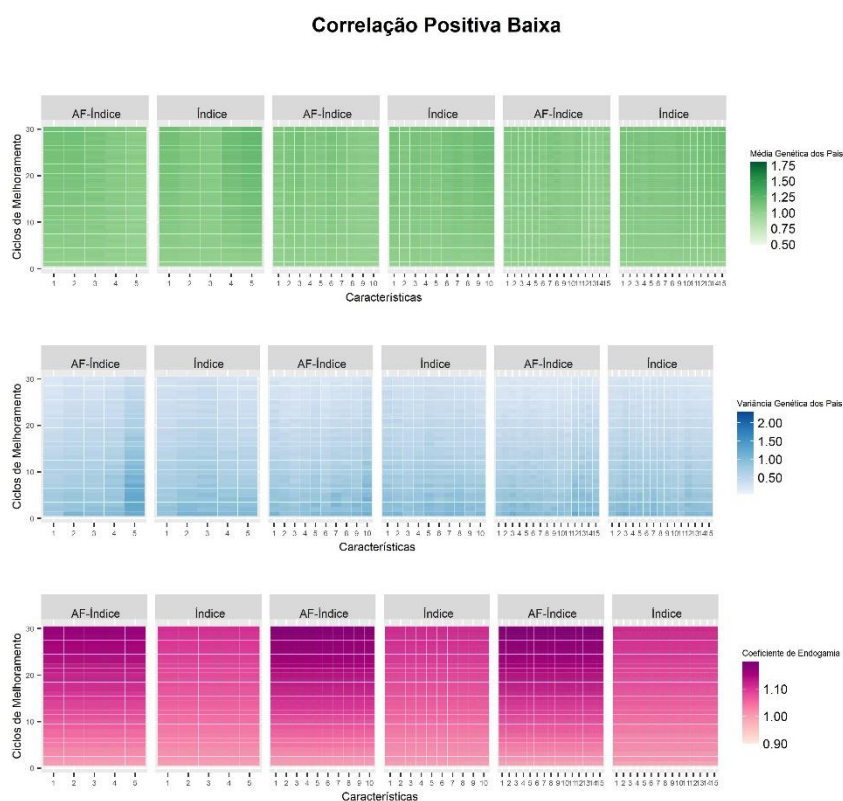


Figura 3. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas são positivamente correlacionadas de baixa magnitude (correlação variando entre 0.00 e 0.30).

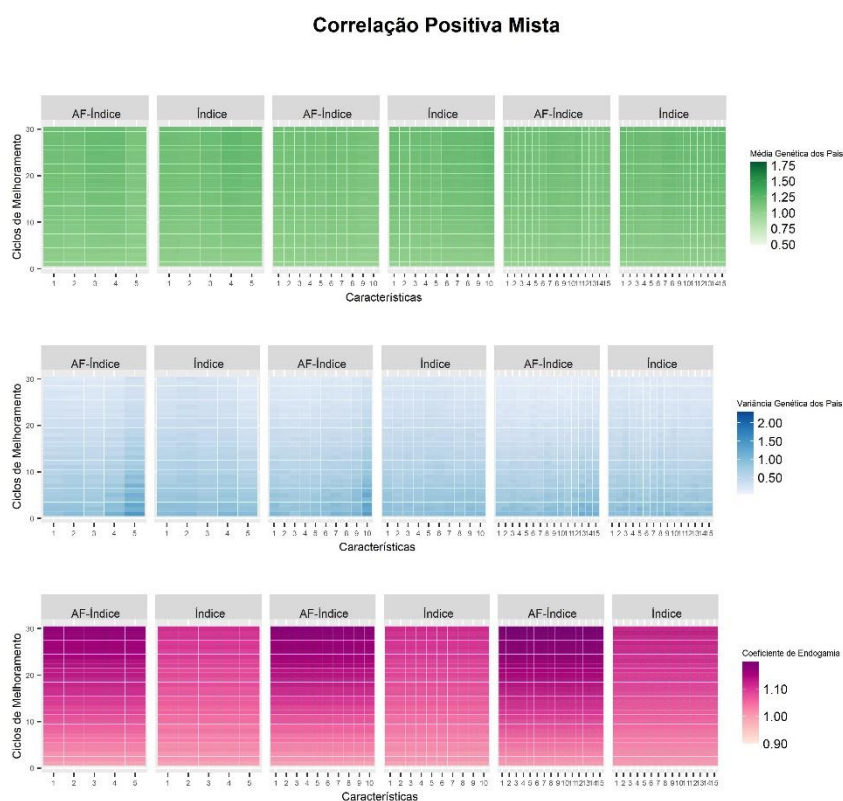


Figura 4. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5$, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações mistas e positivas.

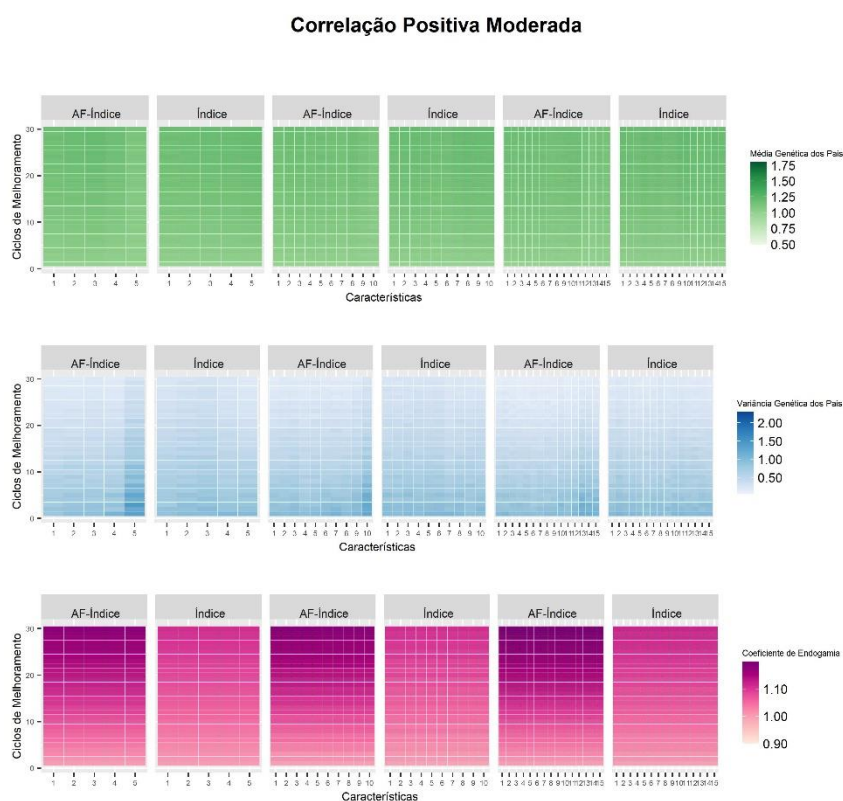


Figura 5. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5, 10$ e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações positivas e moderadas (correlação variando entre 0.40 e 0.60).

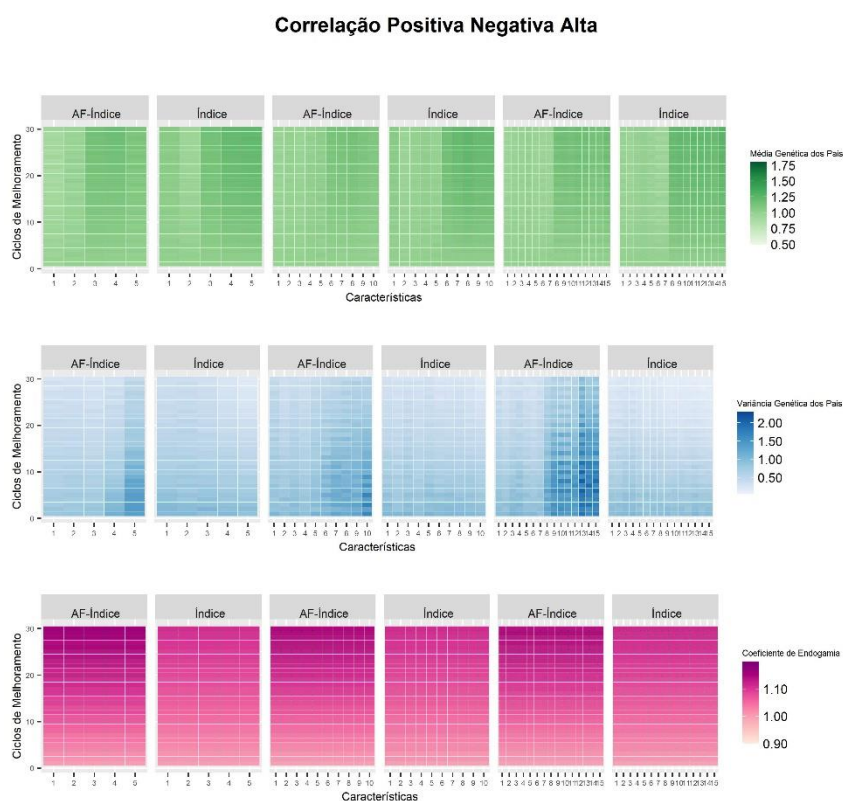


Figura 6. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de Índice e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas são altamente correlacionadas positiva e negativamente (correlação variando de 0.70 a 1 e -1 a -0.70).

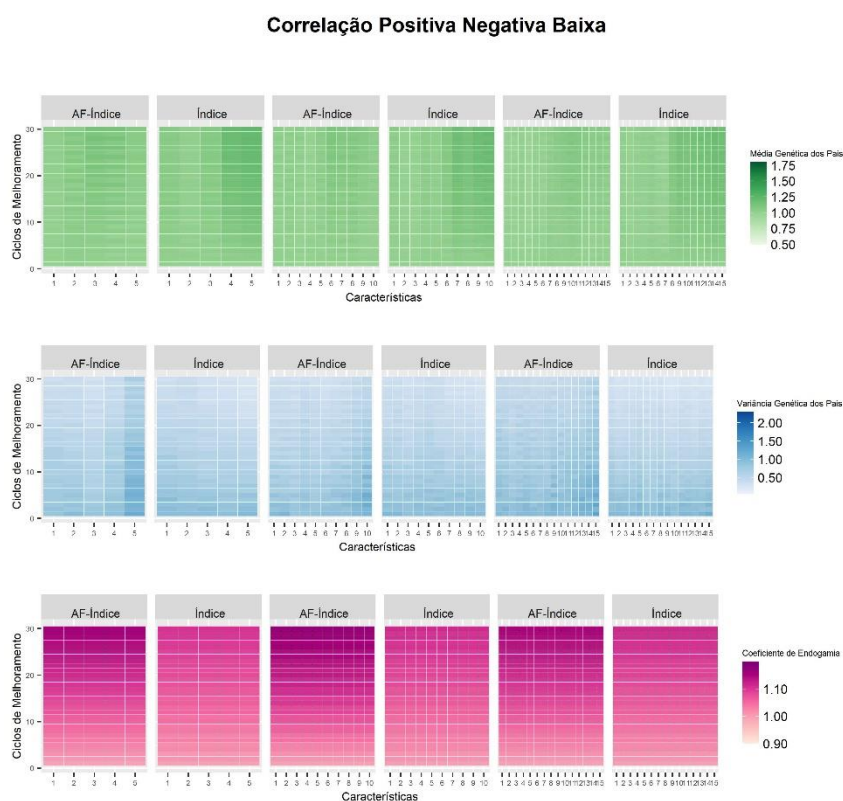


Figura 7. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de Índice e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações positivas e negativas de baixa magnitude (correlação variando de 0.00 a 0.30 e -0.30 a -0.00).

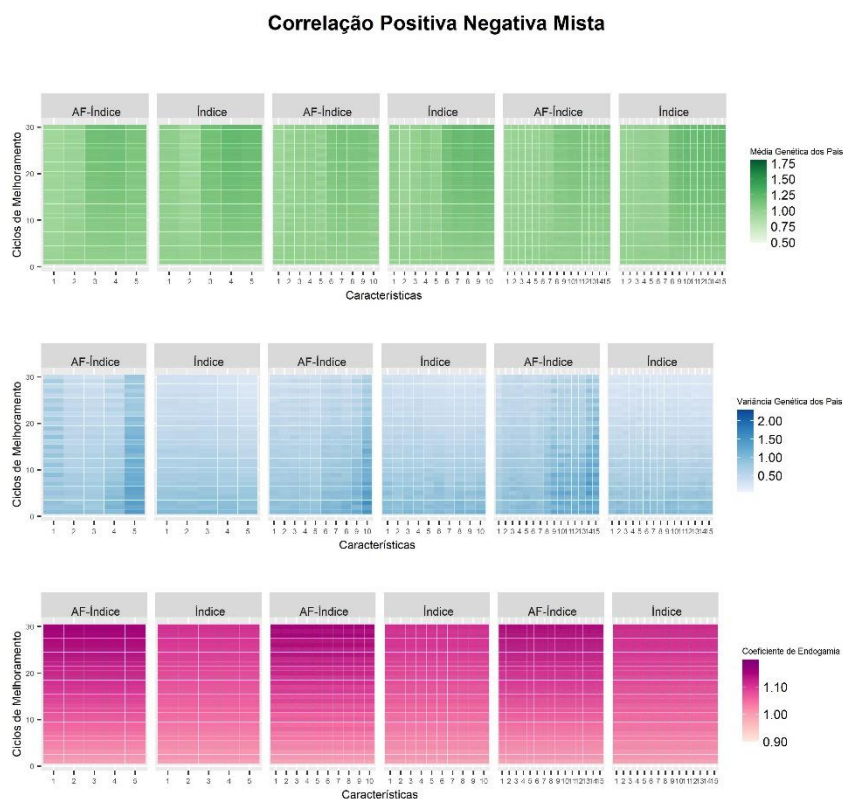


Figura 8. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5$, 10 e 15 características fenotípicas utilizadas nas abordagens de Índice e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações mistas, tanto positivas quanto negativas.

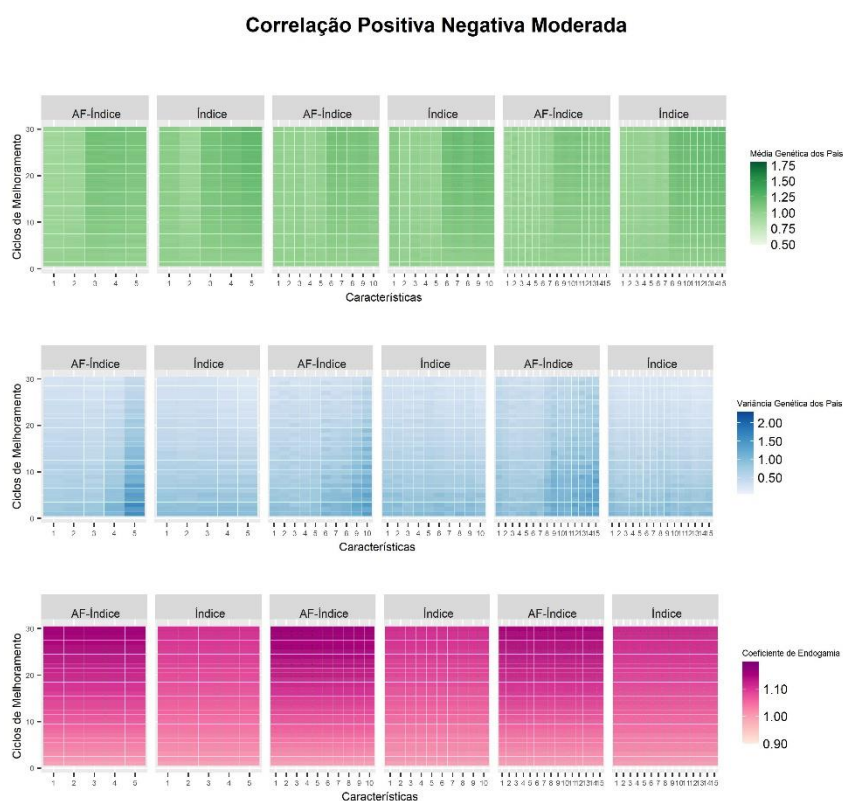


Figura 9. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações positivas e negativas de magnitude moderada (correlações variando de 0.40 a 0.60 e -0.60 a -0.40).

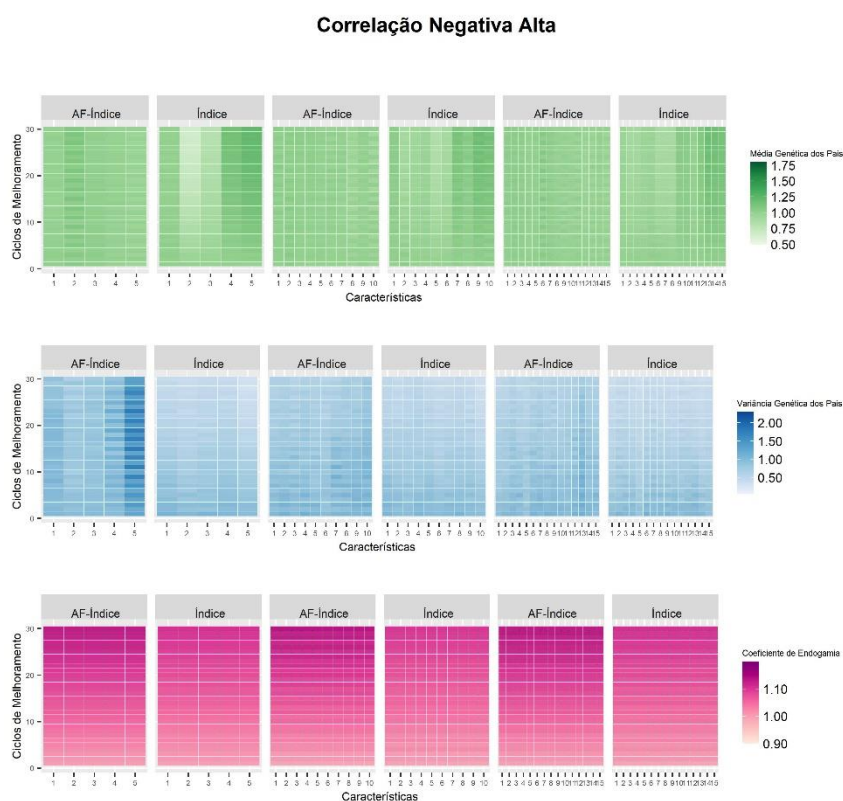


Figura 10. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas são altamente correlacionadas negativamente (correlação variando entre -1.00 e -0.70).

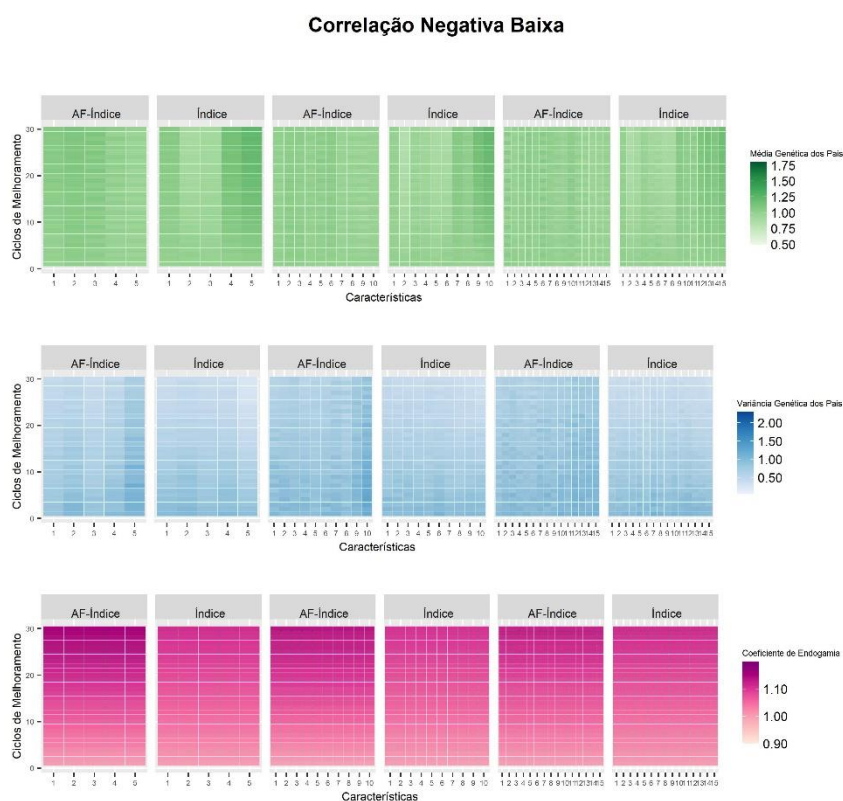


Figura 11. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5$, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações negativas de baixa magnitude (correlação variando entre -0.00 e -0.30).

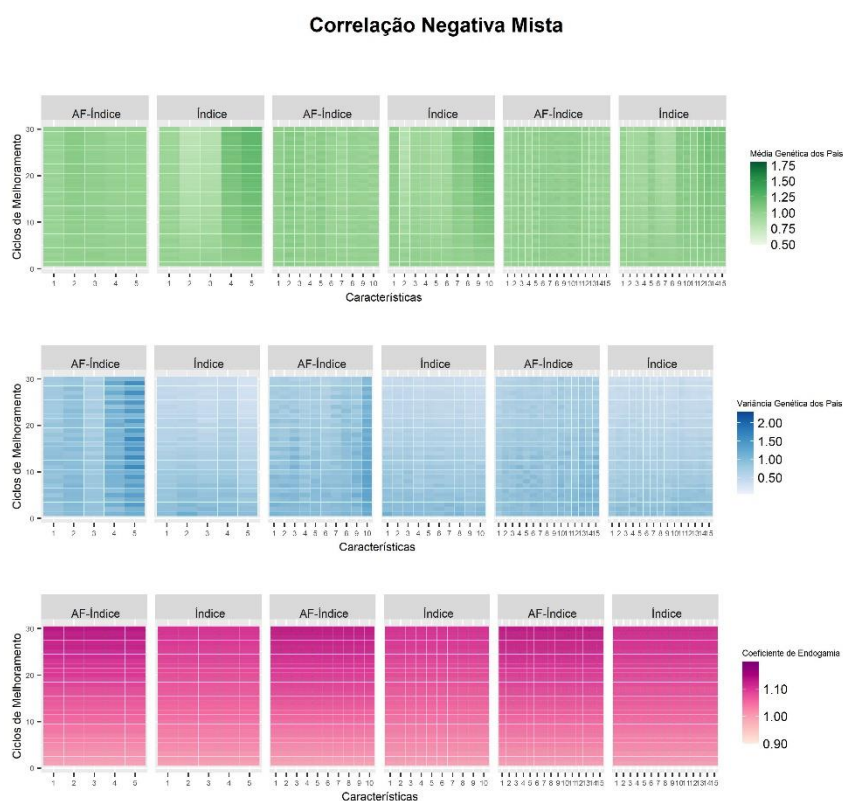


Figura 12. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando $N=5$, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações negativas e mistas.

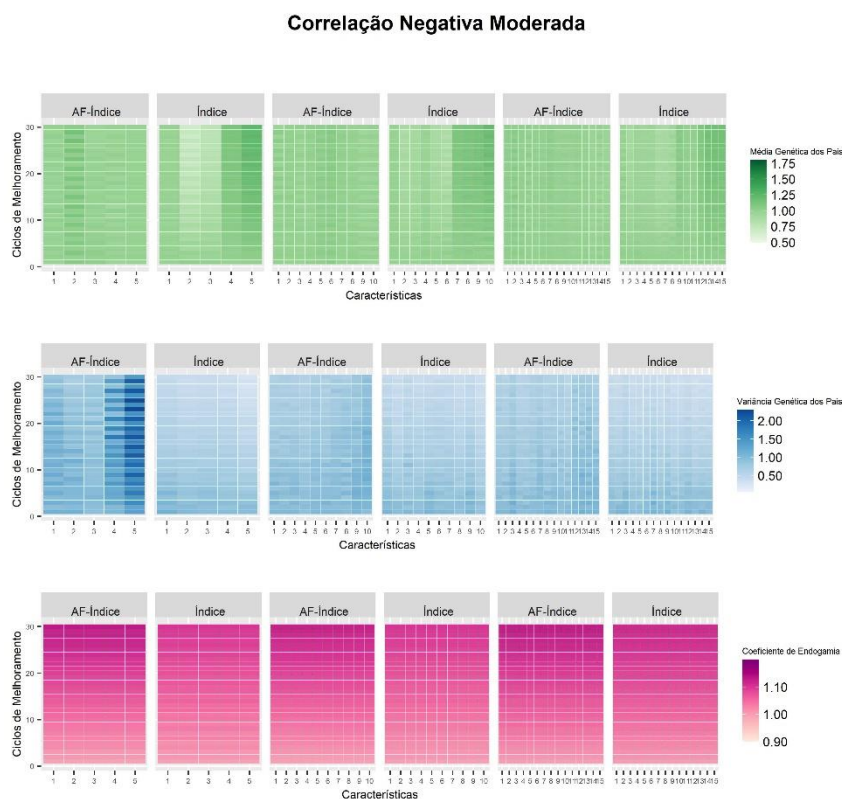


Figura 13. Heatmap contendo o ganho genético (Média Genética dos Pais) em verde, a diversidade genética (Variância Genética dos Pais) em azul e Coeficiente de endogamia em rosa ao longo dos ciclos de melhoramento, considerando N=5, 10 e 15 características fenotípicas utilizadas nas abordagens de índices (Índice) e Análise de Fatores combinada ao índice de seleção (AF-Índice) no cenário em que as variáveis simuladas apresentam correlações negativas e moderadas (correlação variando entre -0.60 e -0.40).

Em relação à magnitude das correlações, destaca-se o cenário correlação altamente positiva em termos de ganho genético (Figura 2). Isso se justifica pelo fato de que estimativas de coeficientes de correlação positiva indicam uma tendência para uma variável aumentar quando a outra também aumenta, porém, neste caso, o alvo dos melhoristas deve ser sempre aumentar todas as características (Leite et al., 2016). Portanto, uma estratégia eficaz para potencializar os ganhos de uma característica específica é aumentar o ganho genético de outra característica altamente correlacionada positivamente com esta. Leite et al. (2016) demonstraram uma alta correlação entre a produtividade de grãos em linhagens de soja e as

características número de nós e número de vagens, indicando que o aumento dessas variáveis resulta em maior produtividade.

Os cenários contendo correlações positivas e negativas (Figuras 6, 7, 8 e 9) apresentaram semelhanças em relação ao ganho genético, com uma tendência de maiores valores para metade das variáveis à medida que avançam os ciclos de seleção. Isso se justifica pela forma como as variáveis foram construídas na simulação, uma vez que as variáveis da segunda metade em diante estão correlacionadas positivamente. Nos quatro cenários com correlação negativa entre as características (Figuras 10, 11, 12 e 13), observa-se um padrão de baixos valores de ganho genético ao longo dos 30 ciclos de melhoramento. Entre esses cenários, destaca-se pelo melhor desempenho em termos de ganho genético aquele com correlação de menor magnitude. Isso ocorre porque este cenário apresenta a menor correlação negativa entre as variáveis, consequentemente, à medida que uma variável aumenta seu ganho genético, a outra diminui, mas com menor intensidade.

Na comparação das duas abordagens de seleção em cada cenário, nos casos de correlação positiva (Figuras 2, 3, 4 e 5), o ganho genético foi equivalente entre as metodologias, reforçando que a combinação da análise fatorial com índices poder ser uma solução viável ao uso exclusivo do índice. Nos cenários de correlação negativa (Figuras 10, 11, 12 e 13), embora o uso exclusivo do índice possa gerar ganhos superiores em algumas características e inferiores em outras, a combinação da AF e índices resulta em ganhos mais homogêneos para todas as características. Essa disparidade nos ganhos genéticos entre as características evidencia o impacto da correlação negativa, em que algumas características apresentam ganhos genéticos maiores, influenciando outras a contabilizar ganhos menores.

2.4.2. Diversidade genética

A diversidade genética é essencial para a adaptação e sobrevivência das populações, permitindo que indivíduos com características genéticas favoráveis persistam em condições adversas (Cruz et al., 2011). A variância genética dos pais indica a amplitude da diversidade genética presente na população ao longo dos ciclos de melhoramento. A ausência ou diminuição significativa da variância genética pode comprometer severamente a adaptabilidade da população a novos desafios, como doenças, alterações climáticas e variações ambientais. Devido aos sucessivos ciclos de melhoramento empregados na seleção recorrente, a variabilidade genética tende

a diminuir (Geraldi, 1997). No entanto, métodos de melhoramento que mantenham a diversidade ao longo dos ciclos são preferíveis.

A avaliação da diversidade genética ao longo dos ciclos revelou uma redução em todos os cenários (correlação positiva, negativa e positiva-negativa), especialmente nos cenários de correlação positiva (Figuras 2, 3, 4 e 5). No subcenário de correlação positiva de alta magnitude, a diversidade genética foi a menor (Figura 2). Este resultado pode ser justificado pela alta correlação positiva entre os caracteres, resultando na escolha de muitos indivíduos geneticamente semelhantes, o que reduz a variância genética. O estudo de Reis et al. (2011) sobre a diversidade genética em maracajuzeiro-amarelo em um programa de seleção recorrente corrobora os resultados, mostrando uma redução na variabilidade após dois ciclos de melhoramento. Diante deste cenário, estratégias para conter a redução da diversidade genética devem ser implementadas, conforme demonstrado nos trabalhos de Lima et al. (2014), que propuseram incorporar 10 progênies com maior representatividade genética no próximo ciclo de recombinação, e de Carias et al. (2016), que sugeriram conduzir simultaneamente a seleção recorrente com a seleção clonal para evitar uma redução significativa na variabilidade genética. No entanto, nos estudos de Moraes et al. (2017) sobre a variabilidade genética em arroz na seleção recorrente, a variabilidade foi mantida durante os quatro primeiros ciclos.

A preservação da diversidade genética para algumas características ocorreu em cenários com correlações mistas e negativas de magnitudes moderada a alta (Figuras 6, 8, 9, 10, 12 e 13). Isso se deve ao fato de que os sinais opostos dessas correlações evitam a seleção de indivíduos geneticamente idênticos, promovendo maior diversidade. Essa preservação foi alcançada por meio da combinação das técnicas de seleção (AF-Índice), o que sugere que essa abordagem pode ser uma estratégia eficaz para conter a redução da variância genética nos primeiros ciclos de melhoramento. Nos cenários de baixa magnitude (Figuras 7 e 11) e correlação exclusivamente positiva (Figuras 2, 3, 4 e 5), ambos os métodos de seleção resultaram em variância genética semelhante.

Nos cenários de correlação mista e negativa de magnitudes mista, moderada e alta (Figuras 6, 8, 9, 10, 12 e 13), que utilizam a combinação das técnicas (AF-Índice) para a seleção dos pais, há uma preservação moderada a alta da diversidade genética para algumas características. Isso ocorre porque os sinais opostos previnem a seleção de indivíduos geneticamente idênticos. Nos cenários de baixa magnitude

(Figuras 7 e 11) e correlação exclusivamente positiva (Figuras 2, 3, 4 e 5), a variância genética é semelhante com ambos os métodos de seleção.

2.4.3. Coeficiente de Endogamia

O coeficiente de endogamia demonstra uma tendência crescente ao longo dos ciclos de seleção em todos os cenários, influenciada pelo padrão de escolha dos progenitores para os ciclos subsequentes. Isso ocorre porque a seleção recorrente favorece a reprodução de indivíduos geneticamente similares para o aumento de alelos favoráveis na população original (Carena et al., 2010), levando a uma maior probabilidade de ligação dos descendentes aos mesmos progenitores em ciclos subsequentes. Essa prática contribui para o aumento da endogamia ao longo do tempo e consequentemente, para a redução da diversidade genética.

Observando os diversos cenários ao longo dos ciclos de melhoramento, nota-se que o de correlação positiva, em todas as suas magnitudes (Figuras 2, 3, 4 e 5), está associado aos maiores valores do coeficiente de endogamia, seguido do cenário de correlação mista (Figuras 6, 7, 8 e 9). Os subcenários relacionados à magnitude da correlação apresentam uma uniformidade entre si. É importante destacar que o cenário em que as variáveis são correlacionadas positivamente também é o de maior ganho. Esse fenômeno é resultado do aumento na frequência de alelos favoráveis, promovendo a reprodução preferencial dos indivíduos geneticamente similares. O trabalho de Mistro et al. (2019) explica que a seleção recorrente com ganhos iniciais elevados, como ocorre em alguns cenários de correlação positiva e mista, pode acarretar prejuízos aos ganhos genéticos ao longo dos ciclos devido ao baixo tamanho efetivo populacional, que resulta em maior endogamia e menor variabilidade genética.

O cenário correlação negativa (Figuras 10, 11, 12 e 13), em comparação com os demais, apresenta os menores valores do coeficiente de endogamia, com os subcenários semelhantes entre si também. Este fato está relacionado a ganhos relativamente menores se comparado com os outros cenários, resultando em valores de endogamia também reduzidos. Estes resultados estão de acordo com o estudo de Guimarães et al. (2018) sobre o impacto da seleção recorrente no milho de pipoca, que evidenciou o aumento da endogamia ao longo dos nove ciclos de seleção recorrente. O trabalho de Nogueira et al. (2019) para a otimização da seleção em eucaliptos verificou que o aumento na endogamia e a redução da diversidade genética

das gerações posteriores são influenciados pelo baixo tamanho efetivo populacional, uma vez que um número baixo de indivíduos geneticamente divergentes participa da recombinação nos cruzamentos.

Ao comparar as técnicas de seleção, seja pela seleção dos pais do próximo ciclo por índice de seleção ou pela combinação do índice com a análise de fatores, em todos os cenários (correlação positiva, negativa e positiva-negativa) e subcenários (baixa, alta, moderada e mista), o coeficiente de endogamia é maior ao empregar a combinação das técnicas. Embora nos cenários de correlação negativa os valores de endogamia sejam mais elevados para a combinação, ambos os métodos de seleção resultam em valores de endogamia mais próximos em comparação com os outros cenários.

2.4.4. Número de variáveis

Conforme mencionado por Torres Filho et al. (2005), Litz et al. (2016) e Cruz et al. (2012), a seleção para várias características diminui o ganho para cada uma delas separadamente. Além disso, à medida que o número de características aumenta, aumenta a probabilidade da presença de multicolinearidade, o que ocasiona matrizes mal condicionadas e estimativas viesadas de ganhos genético utilizando índices de seleção (Olivoto & Nardino, 2020).

Em relação ao ganho genético, os cenários de correlação exclusivamente positiva (Figuras 2, 3, 4 e 5) não mostraram variação com o aumento do número de características para nenhuma das técnicas de seleção utilizadas, sendo o de alta magnitude (Figura 2) o melhor, independentemente do número de variáveis. Nos cenários de correlação negativa (Figuras 10, 11, 12 e 13), a seleção pelo índice demonstrou maior sensibilidade com o aumento das características, com o cenário de baixa magnitude (Figura 11) se destacando para cinco características, em ambas as técnicas de seleção. Nos cenários de correlação mista (Figuras 6, 7, 8 e 9), a combinação das técnicas apresentou maior ganho ao considerar cinco características, enquanto a seleção baseada no índice manteve ganho semelhante, independentemente do número de variáveis.

A diversidade genética foi o parâmetro mais influenciado pelo número de características, especialmente ao combinar técnicas de seleção, enquanto o uso exclusivo do índice mostrou pouca variação. Essa medida diminuiu com o aumento

do número de características nos cenários de correlação negativa (Figuras 10, 11, 12 e 13) e positiva (Figuras 2, 3, 4 e 5) para ambas as técnicas. No entanto, ao empregar a combinação das técnicas para cinco características, observou-se maiores valores de diversidade genética. No cenário de correlação mista de alta magnitude (Figura 6), houve aumento na diversidade genética com o aumento do número de variáveis. Para maiores valores de variância genética, o cenário mais favorável envolve a utilização de cinco variáveis, empregando a combinação de análise fatorial e construção de índices.

A endogamia manteve-se constante com o aumento das características na seleção apenas pelo índice. Já na seleção pela combinação de análise fatorial e índice, nos cenários de correlação mista (Figuras 6, 7, 8 e 9) e negativa (Figuras 10, 11, 12 e 13), a endogamia diminuiu à medida que a quantidade de características sob seleção aumentou. Entretanto, nos cenários de correlação positiva (Figuras 2, 3, 4 e 5), a endogamia tende a aumentar com o aumento do número de características.

5. CONCLUSÕES

A seleção dos progenitores pela combinação da análise fatorial e índices proporcionou ganho genético semelhante e maior diversidade genética entre as abordagens. A direção e a magnitude das correlações entre características influenciam diretamente o ganho genético. Correlações positivas impulsionam ganhos simultâneos, enquanto as negativas reduzem o ganho geral. A magnitude intensifica esse efeito: quanto maior a correlação positiva, maior o ganho; quanto mais acentuada a negativa, maior a redução. Considerando a diversidade genética, cenários de correlações positivas, que geram maiores ganhos, aceleram a redução da diversidade genética. A magnitude da correlação pode intensificar esse efeito, enquanto a endogamia é influenciada pela direção das correlações, mas é pouco sensível à sua magnitude.

A combinação da análise fatorial e dos índices de seleção é sensível à direção e magnitude das correlações. Correlações positivas agrupam variáveis que aumentam juntas, enquanto correlações negativas revelam compensações entre variáveis. A magnitude das correlações influencia a clareza dos fatores: correlações fortes facilitam a distinção dos fatores, enquanto correlações fracas dificultam essa identificação.

Por fim, os cenários com cinco características apresentaram, de maneira geral, melhor desempenho em ganho genético e diversidade para ambas as técnicas utilizadas. Portanto, a combinação da análise fatorial e índices de seleção pode substituir o uso exclusivo do índice na seleção a longo prazo, dependendo dos objetivos de cada estudo. É essencial analisar a correlação entre os caracteres e as técnicas de seleção, além de adotar estratégias para conter a diminuição da diversidade genética e o aumento do coeficiente de endogamia na seleção recorrente a longo prazo, visando obter um equilíbrio ótimo entre ganho genético, diversidade genética e endogamia.

REFERÊNCIAS

- Azevedo, C. F., Ferrão, L. F. V., Benevenuto, J., de Resende, M. D. V., Nascimento, M., Nascimento, A. C. C., & Munoz, P. R. (2024). Using visual scores for genomic prediction of complex traits in breeding programs. **Theoretical and Applied Genetics**, 137(1). <https://doi.org/10.1007/s00122-023-04512-w>
- Bajgain, P., Zhang, X., & Anderson, J. A. (2019). Genome-wide association study of yield component traits in intermediate wheatgrass and implications in genomic selection and breeding. **G3: Genes, Genomes, Genetics**, 9(8), 2429–2439. <https://doi.org/10.1534/g3.119.400073>
- Batista, L. G., Gaynor, R. C., Margarido, G. R. A., Byrne, T., Amer, P., Gorjanc, G., & Hickey, J. M. (2021). Long-term comparison between index selection and optimal independent culling in plant breeding programs with genomic prediction. **PLoS ONE**, 16(5 May 2021). <https://doi.org/10.1371/journal.pone.0235554>
- Borém, A., Miranda, G. V., & Fritsche-Neto, R. (2021). **Melhoramento de plantas**. Oficina de Textos.
- Cabral, P. D. S., do Amaral, A. T., Freitas, I. L. de J., Ribeiro, R. M., & Silva, T. R. da C. (2016). Relação causa e efeito de caracteres quantitativos sobre a capacidade de expansão do grão em milho-pipoca. **Revista Ciencia Agronomica**, 47(1), 108–117. <https://doi.org/10.5935/1806-6690.20160013>
- Campana, A. C. M., Ribeiro Júnior, J. I., & Nascimento, M. (2010). Uma proposta de transformação de dados para a análise de componentes principais. **Revista Brasileira de Biometria**, 28(2), 103–115.
- Carena, M. J., Hallauer, A. R., Miranda Filho, J. B., Hallauer, A. R., Carena, M. J., & Filho, J. B. M. (2010). Selection: Theory. **Quantitative Genetics in Maize Breeding**, 223–290.
- Carias, C. M. de O. M., Gravina, G. do A., Ferrão, M. A. G., Fonseca, A. F. A. da, Ferrão, R. G., Vivas, M., & Viana, A. P. (2016). *Predição de ganhos genéticos via modelos mistos em progênies de café conilon*. **Coffee Sci**. 2016, 11, 39–45.
- Cruz, C. D., & Carneiro, P. C. S. (2003). **Modelos biométricos aplicados ao melhoramento de plantas**. Viçosa: UFV, 2.
- Cruz, C. D., Ferreira, F. M., & Pessoni, L. A. (2011). **Biometria aplicada ao estudo da diversidade genética**. Suprema.
- Cruz, C. D., & Regazzi, A. J. (1994). **Modelos biométricos aplicados ao melhoramento genético**. UFV, Viçosa, 390.
- Cruz, C. D., Regazzi, A. J., & Carneiro, P. C. S. (2012). **Modelos biométricos aplicados ao melhoramento**. UFV, Viçosa.

Entringer, G. C., Santos, P. H. A. D., Vettorazzi, J. C. F., Cunha, K. S. da, & Pereira, M. G. (2014). Correlação e análise de trilha para componentes de produção de milho superdoce. **Revista Ceres**, 61, 356–361.

Falconer, D. S. (1964). **An introduction to quantitative genetics**. Oliver and Boyd, Edinburgh, Second Edition, 312–324.

Faux, A., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., & Hickey, J. M. (2016). AlphaSim: Software for Breeding Program Simulation. **The Plant Genome**, 9(3). <https://doi.org/10.3835/plantgenome2016.02.0013>

Ferreira, A., Roberto Cecon, P., Damião Cruz, C., Gava Ferrão, R., Flores da Silva, M., Francisco Almeida da Fonseca, A., & Amélia Gava Ferrão, M. (2005). Seleção simultânea de Coffea canephora por meio da combinação de análise de fatores e índices de seleção. **Pesquisa Agropecuária Brasileira**, v.40, p.1189-1195.

Ferreira, D. F. (2008). **Estatística multivariada**. Editora Ufla Lavras.

Fialho, I. C., Azevedo, C. F., Nascimento, A. C. C., Teixeira, F. R. F., de Resende, M. D. V., & Nascimento, M. (2023). Factor analysis applied in genomic prediction considering different density marker panels in rice. **Euphytica**, 219(9). <https://doi.org/10.1007/s10681-023-03214-0>

Fritsche-Neto, R., Ali, J., De Asis, E. J., Allahgholipour, M., & Labroo, M. R. (2024). Improving hybrid rice breeding programs via stochastic simulations: number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. **Theoretical and Applied Genetics**, 137(1). <https://doi.org/10.1007/s00122-023-04508-6>

Geraldi, I. O. (1997). *Selección recurrente en el mejoramiento de plantas*.

Guimarães, A. G., Amaral Júnior, A. T. do, Almeida Filho, J. E. de, Pena, G. F., Vettorazzi, C., & Pereira, M. G. (2018). Population structure and impact of recurrent selection on popcorn using EST-SSR markers. **Acta Scientiarum. Agronomy**, 40, e35218.

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. **Genetics**, 28(6), 476–490.

Islam, M. R., Kayess, M. O., Hasanuzzaman, M., Rahman, M. W., Uddin, M. J., & Zaman, M. R. (2016). Selection index for genetic improvement of wheat (Triticum aestivum L.). **Journal of Chemical, Biological and Physical Sciences (JCBPS)**, 7(1), 1.

Jannink, J. L., Astudillo, R., & Frazier, P. (2023). Insight into a two-part plant breeding scheme through Bayesian optimization of budget allocations. **Crop Science**. <https://doi.org/10.1002/csc2.21124>

Leite, W. de S., Pavan, B. E., Matos Filho, C. H. A., Alcantara Neto, F. de, de Oliveira, C. B., & Feitosa, F. S. (2016). Genetic parameters estimation, correlations and selection indexes for six agronomic traits in soybean lines F8. **Comunicata Scientiae**, 7(3), 302–310.

Lima, L. P., Azevedo, C. F., de Resende, M. D. V., Nascimento, M., & Fonseca E Silva, F. (2022). Evaluation of bayesian methods of genomic association via chromosomic regions using simulated data. **Scientia Agricola**, 79(3). <https://doi.org/10.1590/1678-992X-2020-0202>

Lima, V. R., Batista, C. E. A., Bueno, L. G., Resende, T. N., Wendland, A., Pereira, H. S., Faria, L. C., Melo, L. C., & de Souza, T. L. P. O. (2014). Representatividade genética de genitores em progênes de feijão preto do programa de seleção recorrente para tolerância ao mosaico dourado.

Litz, F. H., Castilhano, H., Campos, D. F., Taveira, R. Z., & da Silveira Neto, O. J. (2016). Melhoramento genético de frangos de corte. **Pubvet**, 6, Art-1369. <https://doi.org/10.22256/pubvet.v6n18.1371>

Mistro, J. C., Resende, M. D. V. de, Fazuoli, L. C., & Vencovsky, R. (2019). Effective population size and genetic gain expected in a population of *Coffea canephora*. **Crop Breeding and Applied Biotechnology**, 19, 1–7.

Morais, O. P. de, Melo, P. G. S., Moraes, O. P. de, & Colombari Filho, J. M. (2017). Genetic variability during four cycles of recurrent selection in rice. **Pesquisa Agropecuária Brasileira**, 52(11), 1033–1041.

Nogueira, T. A. P. C., Nunes, A. C. P., Santos, G. A. dos, Takahashi, E. K., Resende, M. D. V. de, & Corradi, I. S. (2019). *Genetic evaluation of eucalyptus full-sib progenies and optimization of selection*. **Scientia Forestalis** 47:451-462

Olivoto, T., & Nardino, M. (2020). MGIDI: A novel multi-trait index for genotype selection in plant breeding. **BioRxiv**, 2020–2027.

Paixão, P. T. M., Nascimento, A. C. C., Nascimento, M., Azevedo, C. F., Oliveira, G. F., da Silva, F. L., & Caixeta, E. T. (2022). Factor analysis applied in genomic selection studies in the breeding of *Coffea canephora*. **Euphytica**, 218(4), 42. <https://doi.org/10.1007/S10681-022-02998-X>

Pesek, J., & Baker, R. J. (1969). Desired improvement in relation to selected indices. **Can. J. Plant. Sci.**, 49, 803–804.

Reis, R. V. dos, Oliveira, E. J. de, Viana, A. P., Pereira, T. N. S., Pereira, M. G., & Silva, M. G. de M. (2011). Genetic diversity in recurrent selection of yellow passion fruit detected by microsatellites markers. **Pesquisa Agropecuária Brasileira**, 46, 51–57.

Resende, M. de, Silva, F. F., & Azevedo, C. F. (2014). **Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção**

genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência. Viçosa: UFV.

Smith, H. F. (1936). A discriminant function for plant selection. **Annals of Eugenics**, 7(3), 240–250.

Team, R. C. (2021). R: A language and environment for statistical computing. URL. **R Foundation for Statistical Computing: Vienna, Austria.**

Teixeira, F. R. F., Nascimento, M., Nascimento, A. C. C., e Silva, F. F., Cruz, C. D., Azevedo, C. F., Paixão, D. M., Barroso, L. M. A., Verardo, L. L., de Resende, M. D. V., Guimarães, S. E. F., & Lopes, P. S. (2016). Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research**, 15(2). <https://doi.org/10.4238/gmr.15028231>

Torres Filho, R. A., Torres, R. A., Lopes, P. S., Pereira, C. S., Euclides, R. F., Araújo, C. V, Silva, M. A., & Breda, F. C. (2005). Estimativas de parâmetros genéticos para características reprodutivas de suínos. **SciELO Brasil**, 57(5), 684–689. <https://www.scielo.br/j/abmvz/a/WrsyvBZX9b7c8mp8YxKPZsz/?lang=pt>

Torres, L. G., Rodrigues, M. C., Lima, N. L., Horta Trindade, T. F., Fonseca e Silva, F., Azevedo, C. F., & DeLima, R. O. (2018). Multi-trait multi-environment Bayesian model reveals g x e interaction for nitrogen use efficiency components in tropical maize. **PLoS ONE**, 13(6). <https://doi.org/10.1371/journal.pone.0199492>

Vachhani, J. H., Raghuwanshi, S. S., Kachadia, V. H., Jivani, L. L., Malav, A. K., Boe=& Bhati, S. S. (2016). Selection indices in groundnut (*Arachis hypogaea* L.). **Electronic Journal of Plant Breeding**, 7(1), 140–144.

Vencovsky, R. (1978). **Effective size of monoecious populations submitted to artificial selection.** <https://www.cabidigitallibrary.org/doi/full/10.5555/19791683752>

Yao, J., Zhao, D., Chen, X., Zhang, Y., & Wang, J. (2018). Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). **The Crop Journal**, 6(4), 353–365.

Yokomizo, G. K.-I., Mochiutti, S., de Queiroz, J. A. L., dos Santos, G. R., Furtado, R. G., Brandão, A. P., & Colares, I. B. (2016). Parameter estimates for genetic characters of assai palm trees fruits in Amapá State. **Ciência Florestal**, 26(3), 985–993.