



**GABRIEL MENDES VILLELA**

**DO THE UNSELECTED GENOTYPES INFLUENCE THE  
IDENTIFICATION OF THE BEST SOYBEAN LINES WITH  
THE INBREEDING GENERATIONS?**

**LAVRAS - MG  
2021**

**GABRIEL MENDES VILLELA**

**DO THE UNSELECTED GENOTYPES INFLUENCE THE IDENTIFICATION OF  
THE BEST SOYBEAN LINES WITH THE INBREEDING GENERATIONS?**

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

Prof. Dr. Adriano Teodoro Bruzi  
Orientador

**LAVRAS – MG  
2021**

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca  
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Villela, Gabriel Mendes.

Do the unselected genotypes influence the identification of the best soybean lines with the inbreeding generations? / Gabriel Mendes Villela. - 2021.

78 p.

Orientador(a): Adriano Teodoro Bruzi.

Tese (doutorado) - Universidade Federal de Lavras, 2021.

Bibliografia.

1. Melhoramento Genético. 2. Genética Quantitativa. 3. *Glycine max* L. I. Bruzi, Adriano Teodoro. II. Título.

**GABRIEL MENDES VILLELA**

**OS GENÓTIPOS NÃO SELECIONADOS INFLUENCIAM NA  
IDENTIFICAÇÃO DAS MELHORES LINHAGENS EM SOJA COM O AVANÇO  
DAS GERAÇÕES DE ENDOGAMIA?**

**DO THE UNSELECTED GENOTYPES INFLUENCE THE IDENTIFICATION OF  
THE BEST SOYBEAN LINES WITH THE INBREEDING GENERATIONS?**

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

APROVADA em 31 de agosto de 2021.

Prof. Dr. Adriano Teodoro Bruzi	UFLA
Prof. Dr. Magno Antonio Patto Ramalho	UFLA
Prof. Dr. Jose Maria Villela Padua	UFLA
Dr. Eduardo De Souza Lambert	BAYER
Prof. Dr. Jose Baldin Pinheiro	ESALQ

Prof. Dr. Adriano Teodoro Bruzi  
Orientador

**LAVRAS - MG  
2021**

*A Deus, pelas graças recebidas e me guiar.*

**OFEREÇO**

*Aos meus pais Nélio e Lenice.*

*Ao meu irmão Mateus.*

*À Isabela*

**DEDICO**

## **AGRADECIMENTOS**

Agradeço a Deus por sempre abençoar e iluminar meu caminho oferecendo as melhores oportunidades possíveis.

Agradeço aos meus pais Nélio e Lenice pelo apoio incondicional, ensinamentos fundamentais e amor.

À Isabela e família, por todo amor, carinho, dedicação e companheirismo por participar em todos os momentos decisivos.

Ao meu irmão Mateus, pelo companheirismo e união.

À Universidade Federal de Lavras (UFLA) e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pela oportunidade.

Ao Professor Adriano Bruzi, pela excelente orientação, amizade, ensinamentos e exemplo de profissionalismo.

Aos Membros da Banca pelos conhecimentos técnicos, contribuições e disponibilidade.

À CAPES pela concessão da bolsa de estudos. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Aos Professores pelos ensinamentos e por compartilharem experiências fundamentais para a minha formação profissional.

Ao Grupo de Pesquisas Soja pela contribuição, ensinamentos, ajuda e trabalho em equipe.

Agradeço a todos meus amigos da Agronomia, aos amigos da República Sem Porteira, aos amigos de Belo Horizonte e todos que contribuíram nesta trajetória.

À Empresa GDM e todos os colaboradores pela oportunidade profissional, ensinamentos corporativos, conhecimento técnico e espaço para dedicar na conclusão do trabalho.

**MUITO OBRIGADO**

## RESUMO GERAL

Na necessidade de incremento em produtividade de soja, entendem-se que diversos setores do agronegócio estão envolvidos, especialmente aqueles relacionados aos aspectos agronômicos da cultura. Neste caso, em especial as pesquisas em melhoramento genético de plantas possuem uma contribuição relevante no incremento de produtividades da soja, como diversas outras culturas produzidas no país. O objetivo deste estudo foi verificar a influência de genótipos não selecionados no ranqueamento de linhagens superiores de soja por meio de conjuntos de dados fenotípicos de progênies avaliadas em diferentes gerações e a viabilidade do uso de simulação computacional em programas de melhoramento como validação cruzada. A abordagem via modelos mistos na seleção de genótipos superiores se mostrou extremamente vantajosa no avanço das pesquisas em melhoramento genético. Outra estratégia de análise que permitiu avanços em programas de melhoramento de plantas foi a utilização da simulação computacional. Em etapas de avaliação dos programas de melhoramento, os genótipos selecionados possuem muitas informações associados a eles, o que possibilita a adoção das diferentes estratégias de pesquisa para a seleção das melhores linhagens. Dessa forma, a literatura permite embasar diferentes estratégias de análises que permitam a classificação dos genótipos superiores na seleção de linhagens de soja com acurácia e viabilidade do emprego da simulação computacional em programas de melhoramento. A utilização de genótipos selecionados e não selecionados em um programa de melhoramento de soja com o avanço das gerações permite alterar drasticamente o ranking, gerando um viés na seleção das melhores progênies com o avanço das gerações de endogamia. Esse comportamento foi observado tanto nas avaliações de dados de campo quanto no aspecto de simulação de dados (validação cruzada). Sugere-se que, na realização da análise estatística dos dados, sejam incluídas informações prévias das progênies por meio de abordagens que considerem o desequilíbrio dos dados, visto que essa alternativa não acarreta custos adicionais para o programa de melhoramento.

**Palavras-chave:** Parâmetros genéticos e fenotípicos; Simulação Computacional; Dados não balanceados; Genética quantitativa.

## GENERAL ABSTRACT

In the need to increase soybean yield, it is understood that several agribusiness sectors are involved, especially related to the agronomic aspects of the crop. In this case, research on plant genetic breeding has a relevant contribution to increase soybean yield, like several other crops produced in the country. The aim of the study was to verify the influence of these unselected genotypes in the ranking of superior soybean lines through phenotypic data sets of progenies evaluated in different generations and the feasibility of using computer simulation in breeding programs such as cross validation. The approach via mixed models in the selection of superior genotypes proved to be extremely advantageous in advancing research in genetic breeding. Another analysis strategy that complements advances in plant breeding programs was the use of computer simulation. In stages of evaluation of breeding programs, the selected genotypes have a lot of information associated with them, which enables the adoption of different research strategies to select the best lines. Thus, the literature allows to support different analyzes of analyzes that follow the classification of superior genotypes in the selection of soybean lines with accuracy and feasibility of using computer simulation in breeding programs. The use of selected and unselected genotypes in a soybean breeding program with the progress of generations allows drastically to change the ranking, generating a bias in the selection of the best progenies with the advancement of inbreeding generations. This behavior was observed both in the field data evaluations and in the data simulation aspect (cross validation). It is suggested that when performing statistical analysis of the data, previous information of the progenies should be included via approaches that consider the imbalance of data, given that this alternative does not entail additional costs for the breeding program.

**Key words:** Genetics and Phenotypic Parameters; Computer Simulation; Unbalanced data; Quantitative Genetics



## LISTAS DE FIGURAS

- Figure 1 - Frequency distribution of average BLUPs associated with YLD related  $F_{3:4}$ ,  $F_{3:5}$  and  $F_{3:6}$  soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation. ....52
- Figure 2 - Frequency distribution of average BLUPs associated with MAT related  $F_{3:4}$ ,  $F_{3:5}$  and  $F_{3:6}$  soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation ..... 52
- Figure 3 - Cluster of graphs of simulated YLD data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 62
- Figure 4 - Cluster of graphs of simulated YLD data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities ..... 63
- Figure 5 - Cluster of graphs of simulated MAT data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 63
- Figure 6 - Cluster of graphs of simulated MAT data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 64
- Figure 7- Cluster of graphs of simulated MAT data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM),

considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 67

Figure 8 - Cluster of graphs of simulated MAT data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 68

Figure 9 - Cluster of graphs of simulated YLD data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 69

Figure 10 - Cluster of graphs of simulated YLD data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities. .... 70

## LISTA DE TABELAS

Table 1 - Characterization of the places evaluated according to the variables altitude (m), latitude (S) and longitude (W).....	43
Table 2 - Description of the experiments (EXP) regarding the crop season (YEAR), inbreeding generations (GENERATION), experimental design (DESIGN), plot size (PLOT) and number of locations (SITES), of replicates (REP), of progenies (PROG) and of checks (CHECK) .....	43
Table 3 - Description of the BLUP's for each scenario and experiments in the phenotypic data analysis .....	45
Table 4 - Heritability coefficients for YLD and MAT traits used in data simulation. ....	47
Table 5 - Summary of the joint analysis of genetic and phenotypic parameters associated with agronomic traits grain yield and full maturity among F <sub>3:4</sub> , F <sub>3:5</sub> and F <sub>3:6</sub> soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation. ....	51
Table 6 - Estimates of the parameters expected gain (EG), realized gain (RG), correlated response (CR) and coincidence index (CI) for YLD and MAT .....	54
Table 7 - Summaries of the estimates of the parameters heridabilitiy ( $h^2$ ), accuracy (rgg'), coefficient of variation ( $CV_e$ ) for the different inbreeding generations advance strategies and coincidence index (CI) and realized gain (GR) at the levels of intensities of selection for the three possible scenarios considering YLD and MAT. ....	58

## SUMÁRIO

	<b>CAPÍTULO 1</b> .....	<b>14</b>
<b>1</b>	<b>INTRODUÇÃO GERAL</b> .....	<b>14</b>
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> .....	<b>16</b>
<b>2.1</b>	<b>Melhoramento da soja (<i>Glycine max</i> L.) no Brasil</b> .....	<b>16</b>
<b>2.2</b>	<b>Abordagem de modelos mistos na seleção de genótipos de soja</b> .....	<b>19</b>
<b>2.3</b>	<b>Simulação computacional no melhoramento de plantas</b> .....	<b>24</b>
	<b>REFERÊNCIAS</b> .....	<b>32</b>
	<b>CHAPTER 2 SEQUENTIAL ANALYSIS CONSIDERING PHENOTYPIC AND COMPUTER SIMULATION DATA IN SOYBEAN BREEDING PROGRAM</b> .....	<b>38</b>
<b>1</b>	<b>INTRODUCTION</b> .....	<b>40</b>
<b>2</b>	<b>MATERIALS AND METHODS</b> .....	<b>42</b>
<b>2.1</b>	<b>Phenotypic data set</b> .....	<b>42</b>
<b>2.2</b>	<b>Statistical analysis of data</b> .....	<b>44</b>
<b>2.2.1</b>	<b>Phenotypic data analysis</b> .....	<b>44</b>
<b>2.2.2</b>	<b>Simulation data analysis</b> .....	<b>47</b>
<b>3</b>	<b>RESULTS AND DISCUSSION</b> .....	<b>49</b>
<b>3.1</b>	<b>Phenotypic data analysis</b> .....	<b>49</b>
<b>3.2</b>	<b>Computer simulation data analysis</b> .....	<b>60</b>
<b>4</b>	<b>CONCLUSIONS</b> .....	<b>72</b>
	<b>REFERENCES</b> .....	<b>73</b>



## CAPÍTULO 1 REVISÃO TEÓRICA MELHORAMENTO GENÉTICO DE SOJA E SIMULAÇÃO COMPUTACIONAL

### 1 INTRODUÇÃO GERAL

Considerada a principal commodity brasileira no setor do agronegócio, a soja é uma cultura a qual permite um balanço extremamente positivo no setor agrícola do país e consequentemente no produto interno bruto nacional. Na última safra, no ano agrícola 2020/21, a produtividade média brasileira de soja, girou em torno de 3513 kg.ha<sup>-1</sup> (CONAB, 2021), sendo considerada o principal produto do setor do agronegócio no país. Esses dados posicionam o Brasil como um dos principais países produtores da cultura no mundo e demonstram a importância desta commodity para a balança comercial brasileira.

Na necessidade de incremento em produtividade de soja, entende-se que diversos setores do agronegócio estão envolvidos, especialmente àqueles relacionados aos aspectos agronômicos da cultura. Neste caso, em especial as pesquisas em melhoramento genético de plantas, possuem uma contribuição relevante no incremento da produtividade da soja, como diversas outras culturas produzidas no país. A obtenção de genótipos superiores a cada ano pode ser considerada como um dos principais desafios dos melhoristas de soja e, por isso, algumas abordagens e análise de dados de um programa de melhoramento são fundamentais para que o objetivo final seja alcançado.

A abordagem via modelos mistos na seleção de genótipos superiores se mostrou extremamente vantajosa no avanço das pesquisas em melhoramento genético. Proposta por Henderson e colaboradores (HENDERSON, 1949), a metodologia permitiu incrementos em acurácia sob condições de desbalanceamento no banco de dados, por retornarem predições mais confiáveis do que as obtidas pelo método dos quadrados mínimos (MQM) ou delineamentos não ortogonais, como em séries de experimentos sequenciais, e/ou quando se dispõe de informações de covariância entre efeitos de parentesco via registros genealógicos ou marcadores moleculares (BERNARDO, 2020; BRUZI, 2008; NUNES; RAMALHO; FERREIRA, 2008; PEREIRA *et al.*, 2017; RESENDE *et al.*, 2020).

Outra estratégia de análise que permitiu avanços em programas de melhoramento de plantas foi a utilização da simulação computacional. No melhoramento de plantas, o uso da simulação é justificado quando: as soluções analíticas não existem ou o grau de dificuldade e o número de variáveis envolvidas não permitem a realização de inferências adequadas sobre o problema; pretende-se comparar a eficiência de um novo procedimento ou técnica em relação,

a outros já existentes e até mesmo consagrados e; os métodos que são rotineiramente empregados em algumas etapas de um programa de melhoramento possam ser modificados para propiciar qualquer melhora ou aumento de sua eficiência. O uso histórico da simulação e o desenvolvimento repentino de opções tecnológicas para os programas de melhoramento indicam que, a simulação computacional se faz essencial por desempenhar um papel cada vez mais relevante nas pesquisas futuras, com foco na concepção e otimização de programas de melhoramento de plantas (BORGES DA SILVA; XAVIER; FARIA, 2021; FERREIRA, 2001).

Em etapas de avaliação dos programas de melhoramento, os genótipos selecionados possuem muitas informações associadas a eles, o que possibilita a adoção das diferentes estratégias de pesquisa para a seleção das melhores linhagens. Dessa forma, a literatura permite embasar diferentes estratégias de análises, que permitam a classificação dos genótipos superiores na seleção de linhagens de soja, com acurácia e viabilidade do emprego da simulação computacional em programas de melhoramento.

## 2 REFERENCIAL TEÓRICO

### 2.1 Melhoramento da soja (*Glycine max* L.) no Brasil

O melhoramento genético se iniciou com a domesticação das plantas há aproximadamente 10.000 anos. Apesar de ter sido um processo demorado, foi muito eficaz pela obtenção de plantas via seleção, adaptadas a ambientes completamente diferentes dos quais elas se originaram e, frequentemente, com desempenho muito superior ao obtido no local de origem (RAMALHO *et al.*, 2012).

No Brasil, o melhoramento de plantas começou com a estratégia de introdução pelos imigrantes, no início do século 20 e contribuiu para avanços consideráveis, no entanto, ainda como uma arte (BERNARDO, 2020). O melhoramento como arte persiste até os dias atuais, porém auxiliado de outras ferramentas para tornar o processo mais eficiente (RAMALHO *et al.*, 2012).

Por ser originária do Nordeste da China, latitude em torno de 45° norte, a adaptação da soja às condições brasileiras foi muito limitada. As primeiras introduções ocorreram, na Bahia, em 1882. Após 10 anos, o Instituto Agrônomo de Campinas (IAC) iniciou os primeiros estudos com o melhoramento da cultura, testando cultivares de interesse para uso forrageiro e em rotação de culturas. Já em 1920, iniciaram-se os primeiros trabalhos com hibridação e seleção. Posteriormente, o IAC continuaria seus trabalhos utilizando o método genealógico e de população (bulk), obtendo cultivares como IAC-1 e Santa Rosa (SEDIYAMA, 2015).

O cultivo comercial de grãos dessa leguminosa só começou a ter expressão econômica no país por volta de 1940, no estado do Rio Grande do Sul. Na década de 50, ampliaram-se as pesquisas com soja, aumentando-se significativamente as coleções de cultivares. Foram introduzidos nos EUA, Japão e em várias regiões do Brasil diferentes cultivares e tipos de soja. Vista como excelente alternativa para a sucessão de culturas com o trigo no sul do Brasil, a produção saltou de 206 mil toneladas, em 1960, para 1,056 milhão de toneladas, em 1969. O sucesso da cultura nos anos 60 continuou na década seguinte, permitindo que a soja se consolidasse como a principal cultura do agronegócio brasileiro, passando de 1,5 milhão de toneladas em 1970, para mais de 15 milhões de toneladas em 1979 (EMBRAPA, 2004).

O objetivo do trabalho de melhoramento de soja era desenvolver cultivares com elevada produtividade, altura da planta e inserção das primeiras vagens adequadas à mecanização da lavoura, com resistência ao acamamento e à deiscência natural das vagens, com resistência a



doenças, com boa qualidade de sementes e alto rendimento de óleo e proteína. O resultado dessas pesquisas foi a substituição gradativa de cultivares ultrapassadas (SEDIYAMA; TEIXEIRA; BARROS, 2009).

Até o final da década de 1960, o melhoramento de soja foi realizado especialmente na região sul do país, regiões de latitude parecida à origem. Durante a década de 1970, cultivares obtidas pelo IAC em parceria com a Embrapa, como Doko e IAC-8 foram determinantes para a expansão da soja no Brasil central juntamente com o avanço do plantio direto (SEDIYAMA, 2015).

Muitos foram os fatores que contribuíram para a expansão da soja no Brasil da região Sul para a região central, associada ao desenvolvimento rápido de tecnologias para suprir a demanda externa (APROSOJA, 2014). Alguns destes fatores podem ser citados como principais: subsídio à triticultura compondo a sucessão de culturas, alta no mercado internacional, substituição das gorduras animais por óleos vegetais, facilidades na mecanização da cultura, estabelecimento de uma pesquisa bem desenvolvida sobre a cultura, mas, sobretudo, à obtenção de cultivares de soja insensíveis ao fotoperíodo, originando as cultivares denominadas de Período Juvenil Longo (PJL) (EMBRAPA, 2004).

A partir de pesquisas desenvolvidas pelo IAC, tendo em vista o melhoramento genético e sua ampla variabilidade genética de resposta, foi possível a seleção de plantas com período juvenil longo. A soja é considerada uma planta de dias curtos. Dessa forma, quando a cultura é cultivada em condições de menor latitude, como ocorre em algumas regiões dos Estados Unidos, a planta acumula, rapidamente, as horas de escuro necessárias para o seu florescimento, assim o crescimento vegetativo é reduzido e a produtividade muito baixa. No Brasil, especialmente, no Centro-oeste e Norte, as condições são, ainda, mais díspares. Neste caso foi preciso selecionar linhagens que antes de florescer, tinham bom desenvolvimento vegetativo, o que se denominou período juvenil longo (PJL) (HARTWIG; KIIHL, 1979). O impulso dado pelas cultivares PJL possibilitou a expansão das fronteiras agrícolas no Cerrado, bem como a incorporação de novas regiões produtivas, como o MATOPIBA (Maranhão, Tocantins, Piauí e Bahia) (FREITAS, 2011).

Minas Gerais também foi um estado pioneiro no melhoramento da soja no Brasil. Os primeiros trabalhos de avaliação e recomendação de cultivares introduzidas foram realizados ainda em 1920. A cultura mostrou-se promissora para a região, e a Universidade Federal de Viçosa iniciou suas primeiras pesquisas visando o melhoramento da soja na década de 1960

(SEDIYAMA, 2015). No ano de 1988 foi instituída a Fundação Triângulo de Pesquisa e Desenvolvimento e, entre as diversas cultivares obtidas por essa instituição, a que mais se destacou foi a MG/BR 46 Conquista, lançada em 1995, tornando-se uma das mais cultivadas no Brasil Central (ZITO *et al.*, 2011). Nos últimos anos, a Universidade Federal de Lavras também iniciou seu programa de melhoramento de soja visando obter cultivares precoces e produtivas adaptadas à região sul de Minas Gerais (SILVA *et al.*, 2016; SOARES *et al.*, 2015).

A partir de 1990, todo o cenário do melhoramento genético da soja no Brasil se transformaria com a entrada das multinacionais no país, especialmente com a lei de proteção de cultivares, que motivaram as empresas a investirem no melhoramento de plantas autógamas. Em 1996, a Monsanto (atualmente Bayer) iniciou suas atividades com cultivares convencionais. No entanto, a partir de 2005, a principal evolução nas cultivares de soja foi a transgênica denominada Roundup Ready RR (GIRALDELI, 2018). As primeiras cultivares RR cultivadas comercialmente no Brasil foram procedentes da Argentina. Entre elas, a cultivar Anta foi amplamente cultivada, principalmente em razão de sua precocidade. Mais tarde, outras empresas como GDM, Pioneer, Syngenta, Basf e Bayer também implementaram os programas de melhoramento de soja transgênica. Atualmente, a soja transgênica ocupa mais de 90% das áreas cultivadas (CONAB, 2021).

No ano de 2013 foi lançado pela Monsanto a tecnologia Intacta RR2 PRO, conferindo além da resistência ao Glifosato, resistência às lagartas, devido a introdução do gene cry1Ac (BAYER, 2018). O evento logo teve adesão dos produtores e atualmente diversas empresas tem cultivares em seu portfólio que possuem a tecnologia. Para os próximos anos, as empresas almejam lançar no mercado novas cultivares que associem piramidação de proteínas, além de resistência à novos herbicidas (CTNBIO, 2017).

No início de 2016, foi lançada através de uma parceria entre BASF e Embrapa, a primeira cultivar de soja com a tecnologia Cultivance, que confere à soja resistência aos herbicidas do grupo químico das imidazolinonas, inibidores da enzima ALS (TREU, 2015). Ainda para o ano agrícola 16/17, a Bayer Cropscience disponibilizou em 11 cultivares a tecnologia Liberty Link™. Tal tecnologia, já comercializada em países como EUA e Canadá, confere resistência ao herbicida Glufosinato de Amônio (SALOMÃO, 2016).

Uma nova tecnologia que também se propõe a dar auxílio ao controle de plantas daninhas resistentes ao glifosato e ao manejo de pragas é a tecnologia Conkesta™ Enlist E3™, a qual vem sendo desenvolvida pela Corteva Agriscience. Trata-se de uma cultivar de soja que

além da tolerância aos herbicidas 2,4-D, glifosato e glufosinato, conta com a biotecnologia Bt de maior espectro de controle de lagartas, com perspectiva de lançamento em 2021 (CORTEVA, 2019).

No ano de 2021, a empresa Bayer lançou comercialmente a tecnologia INTACTA 2 XTEND® na soja. Esta corresponde à terceira geração de transgênicos trazida pela Bayer ao mercado nacional. Tolerante ao herbicida glifosato, a plataforma trará também uma nova ferramenta para o controle de plantas daninhas: o Dicamba, que oferecerá um amplo controle de plantas daninhas de folhas largas, além de proteção adicional às pragas alvo como *Spodoptera cosmioides* e *Helicoverpa armígera* (BAYER, 2021).

O desenvolvimento de novas tecnologias aliado à eficiência de programas de melhoramento de empresas multinacionais, propiciaram ganhos genéticos significativos em rendimento, em diferentes partes do mundo. O ganho genético varia de país para país, com valores entre 9,97 kg ha<sup>-1</sup> (WANG *et al.*, 2016) e 43 kg ha<sup>-1</sup> (DE FELIPE; GERDE; ROTUNDO, 2016). No Brasil, Todeschini *et al.* (2019) encontrou um ganho genético médio de 39,42 kg ha<sup>-1</sup> na região Sul do país, para genótipos lançados entre 1965 e 2011.

No Brasil, a produtividade média da soja tem aumentado, desde meados da década de 1970, a uma taxa média de 42,103 kg ha<sup>-1</sup> ano<sup>-1</sup>e, atualmente, a produtividade média brasileira de soja, gira em torno de 3513 Kg.ha<sup>-1</sup> (CONAB, 2021), sendo considerada o principal produto do setor do agronegócio no país.

Depreende-se que o entendimento de suas características agrônômicas e morfológicas, bem como a resistência e/ou tolerância a estresses bióticos e abióticos, é de fundamental importância para que o seu potencial produtivo seja alcançado. Assim, estudos associando acuradas informações genótípicas e fenotípicas, para predizer o desempenho de progênies e futuros ganhos de produtividade, se faz essencial para que o melhoramento genético da soja no Brasil auxilie a obtenção de cultivares, com característica que permitam rendimentos agrônômicos mais elevados, para cada região agrícola onde se cultiva soja.

## **2.2 Abordagem de modelos mistos na seleção de genótipos de soja**

A abordagem estatística sob modelos mistos foi proposta e desenvolvida por Henderson e colaboradores e descrita em uma série de publicações presentes na literatura (HENDERSON, 1949, 1950, 1963, 1975; HENDERSON *et al.*, 1959).

Devido à necessidade de realização de cálculos mais complicados, a utilização deste método rotineiro de análise se tornou possível, somente a partir do avanço dos recursos computacionais, com o desenvolvimento de softwares estatísticos e algoritmos especializados. Esta estratégia de análise, a princípio, foi largamente utilizada no melhoramento animal, sendo que, a partir da década de 1990, começou a ser utilizada no melhoramento vegetal, em culturas como: soja (CARVALHO; FRITSCH NETO; GERALDI, 2008; HU *et al.*, 2014; PEREIRA *et al.*, 2017; RINCKER *et al.*, 2014; ZHANG *et al.*, 2015), milho (BERNARDO, 1994, 2010; FIGUEIREDO, 2014; VIANA *et al.*, 2011), canola (PIEPHO; MÖHRING, 2006), feijão (BRUZI, 2008; NUNES; RAMALHO; FERREIRA, 2008; RESENDE *et al.*, 2015) e cana-de-açúcar (SILVA *et al.*, 2013).

Modelos mistos, conceitualmente, caracterizam a presença de efeitos fixos e aleatórios no modelo. Para o melhoramento, na prática, os tratamentos genéticos fixos apresentam interesse em estimar médias fenotípicas para fins de ranqueamento e seleção via BLUE (Best Linear Unbiased Estimator). Porém, para tratamentos genéticos aleatórios, o interesse se baseia na estimação dos componentes de variância e covariância dos efeitos aleatórios e ordenamentos dos tratamentos genéticos, para fins de seleção. Estes podem ser realizados via predição de valores genotípicos, por meio da melhor predição linear não viesada (BLUP) e estimativas de componentes de variância, por meio das médias do modelo de máxima verossimilhança restrita (REML), que possibilita um processo de seleção mais acurado (COCKERHAM, 1980; PIEPHO; MÖHRING, 2007; RESENDE *et al.*, 2016).

Porém, optar-se pela conveniência da suposição de um fator como fixo ou aleatório pode estar longe de ser prática inofensiva. Para Eisenhart (1947), efeitos devem ser considerados fixos (constantes) quando o interesse do pesquisador centra-se nos efeitos dos níveis do fator em estudo; caso contrário, os efeitos são considerados aleatórios. Os níveis de um dado fator se constituem numa amostra aleatória de uma população de níveis, estes efeitos devem ser tomados como aleatórios, ao passo que, tendo sido estes níveis previamente escolhidos, estes deverão ser considerados fixos. Para Piepho *et al.* (2008), por exemplo, os genótipos avaliados devem ser considerados como aleatórios pois, representam várias populações que foram submetidas à seleção. A discussão também ocorre na escolha do efeito para blocos em vários trabalhos (MÖHRING; WILLIAMS; PIEPHO, 2015; TOMÉ; BUENO FILHO; FERREIRA, 2002).

Com o desenvolvimento de softwares computacionais, análises baseadas no método de modelos lineares mistos, têm ganhado importância dentre as metodologias para identificação de genótipos superiores. Nesse sentido, o BLUP é considerado um procedimento ideal de avaliação genotípica e de seleção, para efeitos genéticos aditivos, efeitos de dominância e efeitos genotípicos, dependendo da situação (RESENDE *et al.*, 2016). Além disso, o BLUP aumenta a acurácia seletiva e permite o uso simultâneo de inúmeras fontes de informação.

As análises via modelos mistos são especialmente vantajosas sob condições de desbalanceamento no banco de dados, por retornarem predições mais confiáveis do que as obtidas pelo método dos quadrados mínimos (MQM) ou delineamentos não ortogonais, como em séries de experimentos sequenciais, e/ou quando se dispõe de informações de covariância entre efeitos de parentesco, via registros genealógicos ou marcadores moleculares (BERNARDO, 2010; BRUZI, 2008; NUNES; RAMALHO; FERREIRA, 2008; PIEPHO; MÖHRING, 2007).

As análises baseadas em modelos mistos fundamentam - se no modelo misto na forma matricial, apresentado a seguir:

$$y = X\beta + Za + \varepsilon$$

$y$  é o vetor de observações ou fenótipos;  $X$  é a matriz do modelo referente aos efeitos fixos;  $\beta$  é o vetor dos efeitos fixos;  $Z$  é a matriz do modelo referente aos efeitos aleatórios  $a$ ;  $a$  é o vetor dos efeitos aleatórios dos genótipos, com  $a \sim \text{NMV}(0;G)$ ;  $\varepsilon$  é o vetor de erros, com  $\varepsilon \sim \text{NMV}(0; R)$  (HENDERSON, 1975).

Assume-se, para esse modelo que, os efeitos aleatórios de genótipos e dos erros são normalmente distribuídos. Para os erros, tem-se assumido geralmente o pressuposto de independência e homocedasticidade, de modo que a matriz  $R$  é denotada na forma  $R= I$ , sendo  $I$  a matriz identidade da ordem dos dados e a variância ambiental comum. Para genótipos, assume-se a matriz  $G = A$ , sendo que  $G$  é a matriz de covariâncias genéticas dos genótipos, que permite incorporar a informação de parentesco existente, a variância aditiva e a matriz de parentesco genético aditivo entre genótipos. O método de modelos mistos envolve a predição de valores genéticos pelo procedimento da melhor predição linear não viesada e a estimação dos componentes da variância pelo método da máxima verossimilhança restrita (REML) (RESENDE, 2002). Se as matrizes  $G$  e  $R$  forem conhecidas,  $\hat{\beta}$  é chamado melhor estimador linear não viesado (BLUE), e  $\hat{g}$  é denominado de melhor preditor linear não viesado (BLUP).

No melhoramento de soja a aplicação de modelos mistos começou a ser utilizada mais recentemente para se obter sucesso com a seleção de genótipos superiores. Visando avaliar estratégias de escolha de genitores, Panter e Allen (1995) compararam o MQM com o método de modelos mistos via BLUP, utilizando uma série de dados históricos de soja e, em todos os casos analisados, ficou comprovado a superioridade dos cruzamentos obtidos por meio do BLUP, além de menor erro padrão.

A identificação das melhores progênes durante o processo seletivo, pode ser considerada como uma das etapas mais dispendiosas para o melhorista. Por isso, a fim de tornar esse processo menos oneroso, melhorar o ganho com a seleção e a acurácia na predição dos valores genéticos, a aplicação dos modelos mistos tem sido recomendada (BORGES *et al.*, 2010; PINHEIRO *et al.*, 2013; RAMALHO *et al.*, 2012). Em relação à interação genótipo  $\times$  ambiente ( $G \times E$ ), é possível estimar parâmetros genéticos, via máxima verossimilhança restrita (REML) e comparar diferentes estratégias de seleção, no contexto de modelos mistos, para melhoramento da soja. É possível encontrar acurácia seletiva muito alta, indicando alta confiabilidade, identificando genótipos de soja mais produtivos e que apresentam alta adaptabilidade e estabilidade (EVANGELISTA *et al.*, 2021).

Quando se considera o modelo como fixo para estimação das médias dos tratamentos e ordenamento das mesmas, Duarte e Vencovsky (2001), em um trabalho teórico demonstram que, quando se passa de uma análise assumindo genótipos como de efeitos fixos para outra em que tais efeitos são tidos como aleatórios, na ausência de desbalanceamento nos dados, tem-se o efeito ‘*shrinkage*’, que nada mais é do que um encolhimento da distribuição das médias ajustadas dos tratamentos em torno da média geral. Piepho *et al.* (2008) comentam que a propriedade ‘*shrinkage*’ presente no preditor BLUP propicia maior acurácia seletiva quando comparado com o BLUE.

Assim, tem sido recomendado associar a informação de parentesco ao método de modelos mistos, visando tornar o processo seletivo mais eficaz e com resultados mais satisfatórios (PIEPHO *et al.*, 2008). Segundo estes autores, alguns pressupostos devem ser atendidos para se utilizar a informação do parentesco, como, os genótipos avaliados poderem ser rastreados até a mesma população base e esta possuir todos os genótipos não relacionados, assim como, em fase de equilíbrio de ligação ou em equilíbrio de Hardy-Weinberg. Entretanto, esses pressupostos exigem que não ocorra seleção dos genótipos na população base, divergindo em relação à realidade dos programas de melhoramento, em que as populações sofrem intensa

seleção, portanto, essas premissas geralmente não se aplicam fielmente. Porém, mesmo com essas imperfeições, o uso da matriz de parentesco tem apresentado sucesso com a seleção pois, os métodos de verossimilhança levam em conta o viés gerado pelas seleções, em parte porque a matriz de relação genética aditiva  $A$  corrige os padrões de fluxo de informação genética de geração a geração.

A inclusão da informação de parentesco via BLUP para aumentar a precisão na obtenção dos valores genéticos com melhor acurácia e conseqüentemente incrementar os ganhos genéticos com a seleção também foi destacada por Ramalho *et al.* (2012). Em um estudo de simulação esses autores concluíram que, quando a informação de parentesco é considerada, os ganhos genéticos são em geral de maior magnitude do que quando essa informação é ignorada e, por isso, o uso do BLUP utilizando o pedigree no momento da seleção para incrementar os ganhos genéticos é recomendada. Nas sucessivas etapas de avaliação dos programas de melhoramento, alguns genótipos são descartados e outros permanecem, gerando um desbalanceamento. Neste caso, os genótipos selecionados possuem um maior número de informações associados a eles, o que possibilita a adoção da informação de todas as gerações para a seleção das melhores linhagens.

Piepho e Mohring (2006) utilizaram dados da avaliação de cultivares de canola em experimentos de VCU, conduzidos na Alemanha, para comparar a acurácia seletiva na identificação das melhores linhagens, levando em consideração apenas dados da geração de referência, em relação a análise sequencial que considera a performance nas gerações anteriores, adotando-se a abordagem de modelos mistos. Segundo os autores a análise realizada utilizando dados de todas as gerações, propiciou maior precisão em relação à análise considerando apenas a geração de referência.

Considerando o desbalanceamento de dados em um programa de seleção recorrente de feijoeiro visando o aumento da produtividade de grãos, Bruzi (2008) adotou três estratégias de análise para avaliar a produtividade de grãos: a) análise sequencial considerando a combinação das gerações  $S_{0:1}$  a  $S_{0:3}$ ;  $S_{0:1}$  a  $S_{0:4}$ ;  $S_{0:1}$  a  $S_{0:5}$  e todas as progênies; b) análise sequencial envolvendo as progênies comuns nas gerações  $S_{0:1}$  a  $S_{0:5}$  ou  $S_{0:1}$  a  $S_{0:4}$  e c) análise sequencial envolvendo as progênies comuns nas gerações  $S_{0:1}$  a  $S_{0:10}$  ou  $S_{0:1}$  a  $S_{0:8}$ , utilizando o método REML/BLUP. Todas as estratégias foram comparadas com o MQM/ANAVA, tradicionalmente utilizada no programa de seleção recorrente, em que se considera a média da geração anterior. Constatou-se que, a análise realizada tradicionalmente é de baixa eficiência, sobretudo em

gerações iniciais, em comparação ao BLUP, e que a acurácia na seleção das progênes pode ser incrementada, utilizando a análise sequencial por meio do procedimento BLUP.

Resende *et al.* (2015) propuseram o emprego de um índice multigerações via REML/BLUP para aumentar a eficiência do método do bulk dentro de progênes, método esse muito utilizado na condução de progênes segregantes de plantas autógamas, a exemplo da soja. Os autores realizaram estudos de simulação e de campo com a cultura do feijoeiro a fim de comparar o índice multigerações com o procedimento padrão utilizado pelos melhoristas. O índice multigerações se mostrou extremamente vantajoso, por incrementar a acurácia e também o ganho genético por geração com o aumento da endogamia, obtendo-se estimativas de 85% e 12%, respectivamente, em relação ao método tradicional. Segundo os autores, este índice poderia adiantar até uma geração de seleção, além de minimizar o efeito da interação genótipos por ambientes.

Por fim, Pereira *et al.* (2017) estudando a influência do efeito da população na estimativa dos componentes genéticos e fenotípicos das melhores progênes de soja, relataram que uma estimativa mais acurada é possível quando o efeito da população é considerado, possibilitando maiores ganhos com a seleção para produtividade e maturação absoluta. Além disso, o efeito da população apresentou maior razoabilidade de utilização em gerações iniciais de endogamia, geralmente associado a uma maior condição de desbalanceamento do programa.

### **2.3 Simulação computacional no melhoramento de plantas**

A simulação tem sido uma ferramenta eficaz para a avaliação e desenvolvimento de novas estratégias de melhoramento. Em razão do impacto, à complexidade das decisões que devem ser tomadas antes de iniciar o processo de melhoramento e, ao elevado custo de implementar múltiplas opções, a simulação computacional, pode ser um recurso importante. De acordo com Norberto e Dachs (1988) e Ferreira (2001), a simulação computacional consiste em imitar, por meio de programa de computador, o comportamento de um sistema real, para estudar seu funcionamento sob condições alternativas e viabilizar a extração de informações importantes, com baixo custo e rapidez, mediante alterações de interesse nos fatores envolvidos no fenômeno em estudo.

Por isso, a simulação computacional tem sido uma ferramenta eficaz no âmbito da genética e melhoramento de plantas sob vários contextos, como o de populações, do indivíduo



ou do próprio genoma. Ela demanda dos geneticistas o desenvolvimento de modelos biológicos que retratem, da melhor forma possível, os fenômenos de interesse e, dos programadores, as rotinas para o processamento adequado, apesar de impor restrições, para que a influência de certos fatores possa ser avaliada (CRUZ, 2001). Os modelos são inerentemente probabilísticos, por envolverem diversas variáveis de natureza aleatória (FERREIRA, 2001). Estes modelos necessitam ser suficientemente simples para permitir a operacionalização na forma de algoritmos e interpretação de forma adequada. Neste aspecto, surge, em muitos casos, uma desvantagem nos modelos simulados, no que se refere à simplificação exagerada da realidade. Por outro lado, estes modelos permitem a variação nos valores assumidos pelos parâmetros, de tal forma que, uma amplitude realística de fatores pode ser facilmente testada e, mais importante, o efeito relativo destes fatores sobre resultados da seleção pode ser estudado. Por exemplo, um modelo de genoma pode representar a arquitetura genética das espécies de interesse (geração de populações, segregação, recombinação, união de gametas, determinação de fenótipos, seleção, *et al.*) (SUN; PENG; MUMM, 2011; XU, 2010).

Conforme Ferreira (2001), no melhoramento de plantas, o uso da simulação é justificado quando: as soluções analíticas não existem ou o grau de dificuldade e o número de variáveis envolvidas não permitem a realização de inferências adequadas sobre o problema; pretende-se comparar a eficiência de um novo procedimento ou técnica em relação a outros já existentes e até mesmo consagrados e os métodos que são rotineiramente empregados em algumas etapas de um programa de melhoramento possam ser modificados para propiciar qualquer melhora ou aumento de sua eficiência.

Como a maioria dos problemas em genética e melhoramento de plantas, tem natureza aleatória, na simulação de sistemas genéticos, a geração de números aleatórios adquire fundamental importância. Inúmeros programas de computador, linguagens de programação e compiladores têm funções prontas para a geração de números aleatórios. Esses números são gerados em uma sequência que só se repete, após um ciclo superior à ordem dos trilhões de algoritmos. Dessa forma, os números gerados são pseudoaleatórios. Para fins práticos, uma vez que a sequência completa (ciclo) jamais será utilizada, esses números comportam-se de forma genuinamente aleatória (FERREIRA, 2001).

Na utilização de uma técnica de simulação, o pesquisador deve precaver-se contra erros, sejam estes devidos a problemas como, levantamentos amostrais, escolha inadequada das distribuições de probabilidades nos eventos de natureza aleatória, simplificação inadequada da

realidade e erros de implantação do sistema simulado. Para a garantia de sua eficiência pode-se lançar mão de processos de validação. Essa validação consiste em fazer o sistema simulado operar nas condições do sistema real e verificar, por meio de testes de hipóteses e outras análises estatísticas, ou por meio de comparação com situações reais já analisados, se os resultados observados na simulação coincidem com os observados no sistema real (CRUZ, 2001).

Os primeiros indícios de simulação de dados surgiram com a utilização do Método de Monte Carlo, por Von Neuman, em 1940 (MORGAN, 1995). O método fornece soluções aproximadas a uma variedade de problemas matemáticos, que permitem realização de ensaios com amostras de números pseudoaleatórios em um computador. A técnica da simulação de Monte-Carlo se baseia em amostragens aleatórias massivas para obter resultados numéricos, isto é, repetindo sucessivas simulações um elevado número de vezes, o que permite simular a realidade por meio do estudo de uma amostra. Assim, pode-se analisar um número elevado de cenários aleatórios, ampliando a perspectiva de cenários finitos possíveis, realizando análises que se ajustam a uma variabilidade real das variáveis consideradas. O aplicativo dessa técnica baseia-se na identificação das variáveis que se consideram mais significativas, assim como as relações entre elas, para explicar a realidade de um estudo, pela substituição do universo real, por um universo teórico, usando números aleatórios. A única restrição para o uso desse método é a sua impraticabilidade para distribuições cuja função distribuição seja desconhecida ou cuja inversa não seja possível pela não existência de algoritmos numéricos (MELO, 2010; PEÑA, 2001).

A simulação tem tido grande aplicabilidade nas pesquisas de genética e melhoramento de plantas, assim são vários os trabalhos realizados utilizando simulação. Muitos pacotes de software para simulações em melhoramento de plantas estão atualmente disponíveis, e sua disponibilidade visa atender a uma demanda crescente de usuários nas diversas instituições de pesquisa, que manipulam um grande volume de dados, os quais requerem um processamento adequado, para que parâmetros estatísticos e biológicos sejam convenientemente estimados. Esses pacotes têm sido úteis para avaliar estratégias de melhoramento existentes em programas de melhoramento real em campo e têm sido usados para desenvolver novas estratégias de melhoramento genético. Vários pacotes oferecem a simulação de amostras com distribuição normal, a partir de médias e variâncias conhecidas; assim como a simulação de ensaios considerando variáveis com distribuição normal, a partir de parâmetros conhecidos, como média, herdabilidade e coeficiente de variação experimental (SUN; PENG; MUMM, 2011).

Gurgel (2004), também utilizou a simulação visando a obtenção de informações a respeito do número de famílias que devem ser manuseadas, para se ter mais sucesso no melhoramento genético. O autor concluiu que na determinação do número ideal de famílias a serem avaliadas no processo seletivo, deve-se considerar a herdabilidade do caráter; para se obter um ganho genético em torno de 1,10, desvios padrões fenotípicos acima da média na população selecionada, são necessárias 5000 linhagens para uma  $h^2 = 0,30$  e 10 linhagens para  $h^2 = 0,70$ . Outros autores também relataram resultados com relação ao número de progênes/famílias para melhorar a eficiência na seleção via simulação (BERNARDO; CHARCOSSET, 2006; PINTO; LIMA NETO; SOUZA JUNIOR, 2000).

Melo (2010) estudou qual seria o número de genitores para a geração das melhores populações em eucalipto. Foram consideradas seis classes de população base, quatro condições ambientais (herdabilidades distintas) e dois tamanhos de populações sintéticas. Os resultados mostraram que, para a estabilização do ganho esperado com a seleção, para intensidades de seleção de 1% e 5%, utilizando populações base pouco melhoradas, são necessários cinco ou seis genitores, independentemente da herdabilidade e da dominância, e para populações base melhoradas, o número ótimo de genitores deve estar entre seis e quinze para gerar as populações sintéticas.

Para o caso particular de espécies autógamas, as diferentes estratégias de melhoramento distinguem-se, principalmente, pela possibilidade de utilização de variados métodos de condução de populações segregantes. Dessa forma, um grande interesse tem sido focalizado para aperfeiçoar as estratégias em termos de eficiência, para, assim, decidir ou não pela sua adoção.

Fouilloux e Bannerot (1988) realizaram uns dos primeiros estudos por meio de simulação computacional, visando à obtenção de informações a respeito do número ideal de famílias em feijão. Foi mostrado que existe pequena vantagem quando se utiliza um número relativamente grande de famílias, 500, por exemplo, em relação a 50. Embora forneçam informações importantes, especialmente no que se refere à restrição da herdabilidade, estas estão fora da realidade, uma vez que foi considerada uma herdabilidade igual a 100%, o que, para a maioria dos caracteres de importância econômica, não ocorre.

Os métodos bulk-pedigree (BP) e selected bulk method (SB), largamente utilizados pelo CYMMYT, foram comparados via simulação computacional (WANG *et al.*, 2003). As simulações foram realizadas em modelos desenvolvidos para simular o procedimento de

melhoramento para a cultura do trigo. Os métodos foram comparados em termos de ganho genético ajustado, do número de cruzamentos mantidos após um ciclo de melhoramento e da alocação de recursos. Os métodos avaliados neste estudo distinguiram-se apenas pela forma de condução das populações nas gerações segregantes  $F_3$  a  $F_5$ . Após o início, com a seleção das plantas na  $F_2$ , ambos seguiram por três gerações de endogamia pelo método bulk, com a diferença de que, no método SB, apenas as plantas selecionadas foram colhidas em bulk. Após isto, novamente praticou-se a seleção de plantas individuais na geração  $F_6$ . Nas gerações seguintes ( $F_7$  a  $F_{10}$ ), as progênies foram conduzidas por meio de bulk dentro de progênies e avaliadas em experimentos. Os autores concluíram que o ganho genético padronizado com método SB foi de 3,3% superior ao alcançado com BP. Uma maior proporção de cruzamentos foi escolhida pelo método SB, comparado ao método BP. Quanto aos recursos alocados da  $F_1$  a  $F_8$ , o método SB ocupou cerca de 1/3 a menos de área do que o método BP, além de produzir 40% menos progênies durante estas gerações. Assim, Wang *et al.* (2003) concluíram que o método SB foi mais efetivo quanto ao tempo, ao trabalho e aos custos envolvidos, além de resultar em ganho superior.

Nunes *et al.* (2008) realizaram um estudo com o objetivo de avaliar a eficiência do método pedigree, pela inclusão da informação de parentesco entre as progênies, a partir da genealogia e, ainda, compará-lo com o método bulk via simulação computacional e em experimentos de campo. Nas simulações, utilizaram 20 genes segregantes, sob modelo genético puramente aditivo, com frequência alélica de 0,5 e herdabilidade de 10%, 25%, 50% e 75% para seleção com base na média de progênies  $F_{4:5}$ . Foram simulados 1.000 experimentos, para cada configuração e método de condução, referentes à avaliação de 256 progênies  $F_{4:5}$ . Ao comparar a eficiência da seleção considerando a média, o BLUP sem informação do parentesco e o BLUP com informação do parentesco (BLUPA), verificou-se que a seleção com o BLUPA resultou em maior ganho, sobretudo para caracteres de baixa herdabilidade, justificando o trabalho investido para a obtenção do registro genealógico.

Pereira *et al.* (2017) com o objetivo de estudar o emprego do método bulk/ $F_2$ , considerando um grande número de genes segregando e durante algumas gerações, simulou um cruzamento bi-parental obtendo-se 2000 indivíduos na geração  $F_2$ . A partir desses indivíduos foram obtidas 150 progênies  $F_{2:3}$  com 100 indivíduos em cada. Foram realizadas 1000 simulações nas gerações  $F_2$ ,  $F_{2:3}$ ,  $F_{2:5}$  e  $F_{2:8}$  considerando dois tipos de interação alélica: aditiva e dominância completa. A partir dessas simulações obteve-se a porcentagem de locos em

homozigose e heterozigose, assim como a amplitude desses valores, considerando 10, 50, 100 e 200 locos para a população  $F_2$  e 10, 50 e 100 locos para as progênies segregantes. A simulação computacional evidenciou com “n” genes, o que é esperado quando se considera um gene. Desta forma, a variabilidade que será liberada entre os descendentes de uma planta  $F_2$  é semelhante à existente entre plantas da população  $F_2$ . Os autores também concluíram que com um grande número de locos é impossível ter todos os genótipos esperados na descendência ficando evidente que a população  $F_2$  não precisa ser muito grande. Depreende-se também que quando se utiliza o método bulk dentro de  $F_2$ , no final do processo, isto é, quando a frequência de locos em heterozigose tende para zero, tem-se uma mistura de linhagens com a mesma a variabilidade da existente entre progênies e que foi explorada com a seleção.

De acordo com Li *et al.* (2012), as principais aplicações da simulação computacional podem ser divididas em três áreas: comparação do método de melhoramento; mapeamento gênico calculando significância e intervalo de confiança de um QTL; combinação dos modelos da cultura, incluindo arquitetura genética e informações da interação genótipos x ambientes. Por meio da simulação computacional, foi possível estudar os efeitos aditivos e de dominância de um QTL associado a características quantitativas. Foi desenvolvido um método baseado em meta-análise de estimativas previamente estimadas. O modelo de mistura gaussiana do processo de Dirichlet (DPGMM) ajustou o número de componentes e estimou os parâmetros da distribuição genômica. O método foi demonstrado usando vários conjuntos de dados QTL de milho para fornecer estimativas de efeitos aditivos e de dominância para rendimento de grãos e outras características quantitativas. Dessa forma, foi possível representar a arquitetura genética de caracteres, que provavelmente envolvem alguns efeitos grandes, além de muitos efeitos pequenos. A simulação permitiu ser um método robusto com o uso de larga escala de dados multi-alélicos ou com meta-análises de conjuntos de dados QTL menores que podem ser derivados de populações bi parental na estimativa precisa dos parâmetros de distribuição (SUN; MUMM, 2016).

O uso de seleção genômica e simulação foi utilizado em um programa de melhoramento de trigo (*Triticum aestivum* L.) para rendimento e qualidade, para selecionar os melhores cruzamentos ou populações. Neste estudo, os autores compararam cinco modelos (RR-BLUP, BayesA, Bayes B, Bayes Ridge Regression e Bayes LASSO) para seleção genômica (GS) com relação à previsão do cruzamento biparental. Selecionando pais adequados para fazer cruzamentos utilizando a simulação pode interromper a correlação negativa entre características

de produção e qualidade. Um índice de seleção que incorpora rendimento, extensibilidade e máximo de resistência foi sugerido como uma nova característica e sua utilidade para selecionar os cruzamentos com os melhores potencial para melhorar o rendimento e a qualidade simultaneamente foi calculado. Dessa forma foi possível manter a diversidade genética a longo prazo. Concluiu-se que combinar seleção genômica com simulação permite a previsão de desempenho do cruzamento de progênies simuladas e, assim, identificar os pais candidatos antes que os cruzamentos sejam feitos no campo para programas de melhoramento de linha puras (YAO *et al.*, 2018).

A simulação permite também, a modelagem de seleção recorrente fenotípica e seleção genômica sob presença de epistasias nos programas de melhoramento de plantas. Recentemente, foi desenvolvida uma ferramenta de simulação computacional denominada QuMARS, que permitiu comparar as estratégias de seleção recorrentes fenotípica, recorrente associada a marcadores moleculares e seleção genômica em modelos de curto e longo prazo. Para o modelo de seleção considerando marcadores foram considerados o modelo de regressões stepwise e forward. Para seleção genômica, os modelos de predição considerados foram GBLUP, Ridge Regression e regressão por Moore-Penrose. Para gerar genótipos e fenótipos individuais durante a simulação, foram estabelecidos um modelo aditivo e dois modelos genéticos epistáticos considerando três níveis de herdabilidade. Os resultados demonstraram que a seleção respostas de GBLUP baseado em seleção genômica e por marcadores via Forward foram mais consistentes em relação à seleção fenotípica sob o modelo aditivo, principalmente nos primeiros ciclos de seleção. Para modelos epistáticos, os modelos se comportaram de forma contrária. Para os dois modelos de epistasia, variância genética total e o aditivo, o componente de variância aumentou em alguns casos após a seleção. Dessa forma, os autores concluíram que, por meio de simulação, a seleção genômica e fenotípica foram métodos de seleção recorrente eficazes para melhorar reprodução de caracteres quantitativos epistáticas e aditivas (QTL) (ALI *et al.*, 2020).

Por fim, mais recentemente, é possível verificar a utilização intensa da estratégia de simulação computacional no ambiente das “ômicas”. Via simulação é possível entender a similaridade entre locais avaliados em uma escala “ômica” de atributos ambientais, direcionando para a previsão de desempenhos de genótipos não observados. Técnicas recentes de sistema de informações geográficas abriram novas fronteiras para uma melhor compreensão e tratamento da interação genótipos x ambientes e, com base em dados simulados, é possível

uma correspondência precisa de locais com seus genótipos mais apropriados; melhor definição das áreas de reprodução, que têm alta correlação genética, para garantir ganhos de seleção entre os ambientes; e determinação eficiente dos melhores locais para realizar experimentos para análises posteriores. A envirotipagem fornece uma nova classe de marcadores para estudos genéticos, que são razoavelmente baratos, cada vez mais disponíveis e transferíveis entre as espécies (RESENDE *et al.*, 2020).

O uso histórico da simulação e o desenvolvimento repentino de opções tecnológicas para os programas de melhoramento indicam que a simulação computacional se faz essencial, por desempenhar um papel cada vez mais relevante nas pesquisas futuras, com foco na concepção e otimização de programas de melhoramento genético de plantas.

## REFERÊNCIAS

ALI, M. *et al.* Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. **Crop Journal**, [s. l.], v. 8, n. 5, p. 866–877, 2020.

Associação dos Produtores de Soja e Milho do Estado de Mato Grosso - APROSOJA. **A história da soja**. 2014. Disponível em: <https://aprosojabrasil.com.br/2014/sobre-a-soja/a-historia-da-soja/>. Acesso em 22 mar 2020.

BAYER. **Bayer lança Intacta 2 Xtend®, nova geração de soja que irá revolucionar o potencial produtivo do Brasil**. 2021 Disponível em: <https://www.bayer.com.br/pt/midia/bayer-lanca-intacta-2-xtend-nova-geracao-soja-que-ira-revolucionar-potencial-produtivo-brasil>. Acesso em: 22 mar 2021.

BAYER. **Intacta RR2 Pro**. Disponível em: <https://www.intactarr2pro.com.br/intacta-rr2-pro>. Acesso em: 12 dez 2018.

BERNARDO, R. **Breeding for quantitative traits in plants**. 2. ed. Woodbury: Stemma Press, 2010.

BERNARDO, R. **Breeding for quantitative traits in plants**. 3. ed. Woodbury: Stemma Press, 2020.

BERNARDO, R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. **Crop Science**, Madison, v. 34, n. 1, p. 20–25, 1994.

BERNARDO, R.; CHARCOSSET, A. Usefulness of Gene Information in Marker-Assisted Recurrent Selection: A Simulation Appraisal. **Crop Science**, Madison, v. 46, n. 2, p. 614–621, 2006.

BORGES, V. *et al.* Desempenho genotípico de linhagens de arroz de terras altas utilizando metodologia de modelos mistos. **Bragantia**, Campinas, v. 69, n. 4, p. 833–841, 2010.

BORGES DA SILVA, E. D.; XAVIER, A.; FARIA, M. V. Joint modeling of genetics and field variation in plant breeding trials using relationship and different spatial methods: A simulation study of accuracy and bias. **Agronomy**, Basel, v. 11, n. 7, p. 1397, 2021.

BRUZI, A. T. **Aplicações da análise de modelos mistos em programas de seleção recorrente no feijoeiro comum**. 2008. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2008.

CARVALHO, A. D. F.; FRITSCH NETO, R.; GERALDI, I. O. Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 8, n. 3, p. 230–235, 2008.

COCKERHAM, C. C. Random and fixed effects in plant genetics. **Theoretical and Applied Genetics**, Berlin, v. 56, n. 3, p. 119–131, 1980.



COMISSÃO TÉCNICA NACIONAL DE BIOSSEGURANÇA - CTNBIO. Processo de OGM. Brasília, 2017. Disponível em: < <http://ctnbio.mcti.gov.br/processo-de-ogm> >. Acesso em: 08 jul. 2017.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira de Grãos**, Brasília, v. 8, safra 2020/21, n. 12, 2021.

CORTEVA AGRISCIENCE. **Tecnologias**. 2019. Disponível em: < <https://www.corteva.com.br/produtos-e-servicos/tecnologias/sistema-enlist.html> >. Acesso em: 05 mai 2019.

CRUZ, C. D. A informática no melhoramento genético. *In*: NASS, L. L. *et al.* (org.). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT, 2001. p. 1086–1118.

DE FELIPE, M.; GERDE, J. A.; ROTUNDO, J. L. Soybean genetic gain in maturity groups III to V in Argentina from 1980 to 2015. **Crop Science**, Madison, v. 56, n. 6, p. 3066–3077, 2016.

DUARTE, J. B.; VENCOSKY, R. Estimación e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agricola**, Piracicaba, v. 58, n. 1, p. 109–117, 2001.

EISENHART, C. The assumptions underlying the analysis of variance. **Biometrics**, Washington, v. 3, n. 1, p. 1–21, 1947.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **Tecnologias de produção de soja – região central do Brasil - 2005**. Londrina: Embrapa Soja; Embrapa Cerrados; Embrapa Agropecuária Oeste; Fundação Meridional. 2004.

EVANGELISTA, J. S. P. C. *et al.* Environmental stratification and genotype recommendation toward the soybean ideotype: a Bayesian approach. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 21, n. 1, p. 1–8, 2021.

FERREIRA, D. F. Uso de simulação no melhoramento. *In*: NASS, L. L. *et al.* (org.). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT, 2001. p. 1119–1141.

FIGUEIREDO, A. G. **Aplicação de modelos mistos na avaliação de estabilidade e adaptabilidade em milho utilizando dados desbalanceados**. 2014. Dissertação (Mestrado Profissional em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2014.

FOUILLOUX, G.; BANNEROT, H. Selection methods in common bean (*Phaseolus vulgaris* L.). *In*: GEPTS, P. (Ed.). **Genetic resources of *Phaseolus* bean**. Dordrecht: Klüner Academic, 1988. 611 p.

FREITAS, M. C. M. A cultura da soja no Brasil: o crescimento da produção brasileira e o surgimento de uma nova fronteira agrícola. **Enciclopédia Biosfera**, Goiânia, v. 7, n. 12, p. 1–12, 2011.

GIRALDELI, A.L. **Soja RR: Tire suas dúvidas e consiga melhores resultados**. Disponível em: <https://blog.aegro.com.br/soja-rr/>. Acesso em: 12 jun 2018.

GURGEL, F. L. **Simulação computacional no melhoramento genético de plantas**. 2004. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2004.

HARTWIG, E. E.; KIIHL, R. A. S. Identification and utilization of a delayed flowering character in soybeans for short-day conditions. **Field Crops Research**, Amsterdam, v. 2, n. C, p. 145–151, 1979.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, Washington, v. 31, n. 2, p. 423–447, 1975.

HENDERSON, C. R. Estimation of changes in herd environment. **Journal of Dairy Science**, [s. l.], v. 32, n. 8, p. 706–706, 1949.

HENDERSON, C. R. Estimation of Genetic Parameters. **Annals of Mathematical Statistics**, Ann Arbor, v. 21, p. 309–310, 1950.

HENDERSON, C. R. Selection index and expected genetic advance. In: HANSON, W. D.; ROBINSON, H. F. (org.). **Statistical genetics and plant breeding**. Washington: National Academy of Science, 1963. p. 141–163.

HENDERSON, C. R. *et al.* The Estimation of Environmental and Genetic Trends from Records Subject to Culling. **Biometrics**, Washington, v. 15, n. 2, p. 192–218, 1959.

HU, Z. *et al.* Association mapping of yield-related traits and SSR markers in wild soybean (*Glycine soja* Sieb. and Zucc.). **Breeding Science**, Tokyo, v. 63, n. 5, p. 441–449, 2014.

LI, X. *et al.* Computer Simulation in Plant Breeding. In: SPARKS, D. L. (org.). **Advances in Agronomy**. San Diego: Academic Press, 2012. v. 116, p. 219–264.

MELO, D. S. **Número ideal de genitores para a geração de populações sintéticas em eucalipto via simulação**. 2010. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2010.

MÖHRING, J.; WILLIAMS, E.; PIEPHO, H. P. Inter-block information: to recover or not to recover it? **Theoretical and Applied Genetics**, Berlin, v. 128, n. 8, p. 1541–1554, 2015.

MORGAN, B. J. T. **Elements of simulation**. 6. ed. London: Chapman & Hall, 1995.

NORBERTO, J.; DACHS, W. **Estatística computacional: uma introdução em turbo-pascal**. Rio de Janeiro: LTC, 1988.

- NUNES, J. A. R.; RAMALHO, M. A. P.; FERREIRA, D. F. Inclusion of genetic relationship information in the pedigree selection method using mixed models. **Genetics and Molecular Biology**, Ribeirão Preto, v. 31, n. 1, p. 73–78, 2008.
- PANTER, D. M.; ALLEN, F. L. Using best linear unbiased predictions to enhance breeding for yield in soybean: II. Selection of superior crosses from a limited number of yield trials. **Crop Science**, Madison, v. 35, n. 2, p. 405–410, 1995.
- PEÑA, D. Deducción de distribuciones: el método de Monte Carlo. *In*: PEÑA, D. (org.). **Fundamentos de Estadística**. Madrid: Alianza Editorial, 2001. p. 193–198.
- PEREIRA, F. C. *et al.* Implications of the population effect in the selection of soybean progeny. **Plant Breeding**, Berlin, v. 136, n. 5, p. 679–687, 2017.
- PIEPHO, H. P. *et al.* BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, Wageningen, v. 161, n. 1–2, p. 209–228, 2008.
- PIEPHO, H. P.; MÖHRING, J. Computing heritability and selection response from unbalanced plant breeding trials. **Genetics**, Austin, v. 177, n. 3, p. 1881–1888, 2007.
- PIEPHO, H. P.; MÖHRING, J. Selection in cultivar trials—Is it ignorable? **Crop Science**, Madison, v. 46, n. 1, p. 192–201, 2006.
- PINHEIRO, L. C. M. *et al.* Parentesco na seleção para produtividade e teores de óleo e proteína em soja via modelos mistos. **Pesquisa Agropecuária Brasileira**, Brasília, v. 48, n. 9, p. 1246–1253, 2013.
- PINTO, R. M. C.; LIMA NETO, F. P.; SOUZA JUNIOR, C. L. Estimativa do número apropriado de progênies S1 para a seleção recorrente em milho. **Pesquisa Agropecuária Brasileira**, Brasília, v. 35, n. 1, p. 63–73, 2000.
- RAMALHO, M. A. P. *et al.* **Aplicações da genética quantitativa no melhoramento de plantas autógamas**. Lavras: Editora UFLA, 2012.
- RESENDE, M. D.V. *et al.* Selection index with parents, populations, progenies, and generations effects in autogamous plant breeding. **Crop Science**, Madison, v. 56, n. 2, p. 530–546, 2016.
- RESENDE, M.D.V. **Efeitos fixos ou aleatórios de repetições no contexto dos modelos mistos no melhoramento de plantas perenes**. Colombo: EMBRAPA Florestas, 2002.
- RESENDE, M. D.V. *et al.* Multigeneration index in the Within - Progenies bulk method for breeding of Self-Pollinated plants. **Crop Science**, Madison, v. 55, n. 3, p. 1202–1211, 2015.
- RESENDE, R. T. *et al.* Enviromics in breeding : applications and perspectives on envirotypic - assisted selection. **Theoretical and Applied Genetics**, Berlin, v. 134, p. 95–112, 2020.

- RINCKER, K. *et al.* Genetic improvement of U.S. soybean in maturity groups II, III, and IV. **Crop Science**, Madison, v. 54, n. 4, p. 1419–1432, 2014.
- SALOMÃO, R. 2016. Bayer lança comercialmente a soja Liberty Link. **Revista Globo Rural**. Disponível em: <https://revistagloborural.globo.com/Noticias/Empresas-e-Negocios/noticia/2016/03/bayer-lanca-comercialmente-soja-liberty-link.html>. Acesso em: 22 nov 2016.
- SEDIYAMA, T. **Melhoramento genético da soja**. Londrina: Mecenias, 2015.
- SEDIYAMA, T.; TEIXEIRA, R. C.; BARROS, H. B. Cultivares. *In*: SEDIYAMA, Tuneo (org.). **Tecnologias de produção e usos da soja**. Londrina: Mecenias, 2009. p. 77–91.
- SILVA, F. F. *et al.* Bayesian inference of mixed models in quantitative genetics of crop species. **Theoretical and Applied Genetics**, Berlin, v. 126, n. 7, p. 1749–1761, 2013.
- SILVA, K. B. *et al.* Adaptability and phenotypic stability of soybean cultivars for grain yield and oil content. **Genetics and Molecular Research**, Ribeirão Preto, v. 15, n. 2, p. 1–11, 2016.
- SOARES, I. O. *et al.* Adaptability of soybean cultivars in different crop years. **Genetics and Molecular Research**, Ribeirão Preto, v. 14, n. 3, p. 8995–9003, 2015.
- SUN, X.; MUMM, R. H. Method to represent the distribution of QTL additive and dominance effects associated with quantitative traits in computer simulation. **BMC Bioinformatics**, [s. l.], v. 17, n. 1, p. 1–15, 2016.
- SUN, X.; PENG, T.; MUMM, R. H. The role and basics of computer simulation in support of critical decisions in plant breeding. **Molecular Breeding**, Dordrecht, v. 28, n. 4, p. 421–436, 2011.
- TODESCHINI, M. H. *et al.* Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. **Euphytica**, Wageningen, v. 215, n. 7, 2019.
- TOMÉ, M. A.; BUENO FILHO, J. S. S.; FERREIRA, D. F. Blocos fixos ou aleatórios? O caso dos ensaios em látice no melhoramento vegetal. **Ciência e Agrotecnologia**, Lavras, v. 26, n. 3, p. 576–584, 2002.
- TREU, F. BASF e Embrapa lançam Cultivance. **Revista Cultivar**, Pelotas, 25 ago. 2015. Disponível em: <https://revistacultivar.com.br/noticias/basf-e-embrapa-lancam-cultivance>. Acesso em: 22 nov 2016
- VIANA, J. M. S. *et al.* Best linear unbiased prediction and family selection in crop species. **Crop Science**, Madison, v. 51, n. 6, p. 2371–2381, 2011.
- WANG, C. *et al.* Seventy-five years of improvement of yield and agronomic traits of soybean cultivars released in the yellow-huai-hai river valley. **Crop Science**, Madison, v. 56, n. 5, p. 2354–2364, 2016.

WANG, J. *et al.* Comparison of two breeding strategies by computer simulation. **Crop Science**, Madison, v. 43, n. 5, p. 1764–1773, 2003.

XU, Y. **Molecular Plant Breeding**. Wallingford: CABI, 2010.

YAO, J. *et al.* Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). **Crop Journal**, [s. l.], v. 6, n. 4, p. 353–365, 2018.

ZHANG, H. *et al.* Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. **Plant Breeding**, Berlin, v. 134, n. 5, p. 564–572, 2015.

ZITO, R. K. *et al.* Tecnologias para o Cerrado Mineiro. **Informe Agropecuário**, [s. l.], v. 32, p. 116, 2011.

## **CHAPTER 2**

### **SEQUENTIAL ANALYSIS CONSIDERING PHENOTYPIC AND COMPUTER SIMULATION DATA IN SOYBEAN BREEDING PROGRAM**

## ABSTRACT

The goal of this study was to verify the influence of unselected genotypes in the ranking of superior soybean lines using phenotypic data set from progenies evaluated in different generations and the feasibility of using computer simulation in breeding programs as cross validation. We evaluated 196 progenies for grain yield and full maturity during three generations ( $F_{3:4}$ ,  $F_{3:5}$ , and  $F_{3:6}$ ). In the experiment “A”, the effect of selection was considered as the inbreeding generations. Regarding experiment “B”, refers to the planting of all progenies, without the selection effect. The defined strategies were evaluations based on reference generation with selection (BLUP-A), without selection taking into account all genotypes per generation (BLUP-B), and a joint analysis of all genotypes and years (BLUP-J). In parallel, an analysis was carried out via computer simulation using the same reasoning proposed above in order to perform a cross-validation with field data. We generated 100 different biparental populations where parental genotypes were randomly assigned, considering 339 and 325 QTL associated with grain yield and plant maturity, respectively. We distributed QTL along the chromosomes according to their position information (available on SoyBase) and sampled additive effects. We performed 100 simulations for each heritability combination and base population and constructed each population. It was performed successive selfing generations using progeny bulks ( $F_{3:4}$  to  $F_{3:6}$ ) and conducted different selection strategies. The strategies proposed in this case were: 1 - Single Generation ( $F_{3:6}$ ), in which the progenies were advanced considering the selection effect; 2 -  $F_{3:6}$  Full Progeny, disregarding the selection effect; and 3 - Multiple Generation Progenies Mean (MGPM), simulating a joint analysis including year effect. Regarding phenotypic and simulation computational analysis, to quantify the efficiency of the inclusion of the unselected progenies in the model, we estimated genetic expected and realized gain with the selection, the coincidence index, and correlated response of selected genotypes for several selection intensities (5%, 15%, and 25%). The use of selected and unselected genotypes in a soybean breeding program with the progress of generations allows drastically to change the ranking, generating a bias in the selection of the best progenies with the advancement of inbreeding generations. This behavior was observed both in the field data evaluations and in the data simulation aspect (cross validation). It is suggested that when performing statistical analysis of the data, previous information of the progenies should be included via approaches that consider the imbalance of data, given that this alternative does not entail additional costs for the breeding program.

## 1 INTRODUCTION

The breeding of autogamous plants has, over the years, been successful in many cultivated species (GORNICKI; FARIS, 2014; RAMALHO; DIAS; CARVALHO, 2012; RAMALHO; MARQUES; LEMOS, 2021). However, considering that the yield of the lines is currently relatively high, breeders should require new tools to understand the strategies already used, if possible, the optimization of breeding programs.

In the successive generation of evaluation of breeding programs, some genotypes are discarded, and others remain, generating a great experimental unbalance. It appears that breeding companies perform hundreds of hybridizations annually and obtain a thousand lines. At the time of decision on which line(s) will be recommended, they quickly and drastically reduce the number of progenies to be evaluated in the experiments (BERNARDO, 2020). It is known that the selected genotypes have a greater number of data associated with them, which allows the adoption of information from all generations to select the best lines, although this strategy has not been widely used by breeding programs (GLENN *et al.*, 2017).

Yield is also the most complex trait to characterize from both a phenotypic and genotypic perspective. It has complex interaction of many genetic and environmental factors that contribute collectively to the final quantitative measurement. This makes it difficult, time-consuming, and expensive to identify progeny with the highest yield potential.

In order to help breeders, a tool would be, SoyBase, a USDA (United States Department of Agriculture) genetic database, holds professionally curated soybean genetic and genomic data, which is integrated and made accessible to researchers and breeders. The site holds several reference genome assemblies, as well as genetic maps, mapped traits, expression and epigenetic data, pedigree information, and extensive data sets. SoyBase displays include genetic, genomic, and epigenetic maps of the soybean genome (BROWN *et al.*, 2021; SCHMUTZ *et al.*, 2010).

Among the strategies that are extremely advantageous in data analysis in research of a plant breeding program, one of them would be the approach via mixed models. Proposed by Henderson *et al.* (HENDERSON, 1949), the method allowed increments in accuracy under unbalanced conditions in the database by returning more reliable predictions than those obtained by the least-squares method (MQM) or non-orthogonal designs such as series of sequential experiments and/or when information on the covariance between kinship effects is



available via genealogical records or molecular markers (BERNARDO, 2010; BRUZI, 2008; LEMOS; ABREU; RAMALHO, 2020; PEREIRA *et al.*, 2017; PIEPHO *et al.*, 2008).

Another analysis strategy that allowed advances in plant breeding programs was the use of computer simulation. In plant breeding, the use of simulation is justified when: (a) the analytical solutions do not exist or the degree of difficulty and the number of variables involved do not allow adequate inferences to be made about the problem; (b) it is intended to compare and/or validate the efficiency of a new procedure or technique with others that already exist and even established, and the methods that are routinely used can be modified to provide any improvement or increase in their efficiency; (c) and answering questions would be impractical in field research due to the huge number of lines, crosses, populations and generations to be evaluated (BORGES DA SILVA; XAVIER; FARIA, 2021).

Several studies have been carried out and the contributions of simulation to genetics and plant breeding have been important both in classical areas, such as quantitative genetics (ABREU *et al.*, 2010; BORGES DA SILVA; XAVIER; FARIA, 2021) and in newer areas such as genomics (DE LOS CAMPOS; SORENSEN; GIANOLA, 2015). This fact can be seen when evaluating some recent applications to answer questions from the most varied aspects. For example, in determining the number and size of families to be evaluated (FERREIRA *et al.*, 2000; FOUILLOUX; BANNEROT, 1988; LEITE *et al.*, 2009; SILVA; VIEIRA; VILLELA, 2009), in verifying the efficiency of diallels in the choice of parents (VEIGA; FERREIRA; RAMALHO, 2000), in the comparison of breeding strategies (KERR; DIETERS; TIER, 2004; NUNES; RAMALHO; FERREIRA, 2008; WANG *et al.*, 2003), and several other works studying different aspects related to the genetic breeding of plants (GERLOFF; SMITH, 1988; SILVA; FERREIRA; PACHECO, 2000; SOUZA; PETERNELLI; BARBOSA, 2006).

The current study was designed primarily to answer a simple yet important question: Do the unselected genotypes influence the identification of the best lines in the soybean breeding program? It is common for breeders to decide which genotypes will be selected based on yield trials only by reference generation, when at most using information from the previous year. A strategy to optimize the data analysis of the breeding program would be to propose a sequential analysis model which would make it possible to contemplate the year effect and the progenies that were not selected as the generations progressed. It is known that the crop season effect is extremely pronounced in soybean genetic breeding programs, especially in tropical regions, and the possibility of involving all genotypes allows for greater accuracy in decision-making.

Thus, we aim to study the influence of unselected genotypes in the ranking of superior soybean lines using phenotypic data set from progenies evaluated in different generations and the feasibility of using computer simulation for sequential analysis in breeding programs as cross validation.

## **2 MATERIALS AND METHODS**

### **2.1 Phenotypic data set**

Data set were evaluated using segregating progenies  $F_{3:4}$ ,  $F_{3:5}$ , and  $F_{3:6}$ , from the Soybean Breeding Program of the Universidade Federal de Lavras, based on a biparental cross between the cultivars P98Y11 RR x BRS FAVORITA RR. The crossing was carried out in a greenhouse in the 2013/2014 season and the population was advanced in Bulk until  $F_3$ . In the 2015/2016 summer crop, 200 plants from this population were selected for progenies evaluation.

The  $F_{3:4}$ ,  $F_{3:5}$ , and  $F_{3:6}$  progenies were evaluated in experiments carried out during three crop seasons as described in the locations and experiments in Table 1 and 2. In all environments, the experiments were carried out in a lattice incomplete design with two or three replications. Plots were constituted by specific sizes which all spaced at 0.50 meters between plant rows. Borders external to the experiment were used.

It was necessary to denominate two experiments distinct to evaluate the progenies. In the experiment “A” the effect of selection was considered as the inbreeding generations. The selection intensity in the  $F_{3:4}$  generation was 50% (100 progenies) and for the  $F_{3:5}$  generation the index was approximately 25% (25 progenies). Regarding the experiment “B”, which was conducted in parallel in those same crop seasons, refers to the planting of all progenies in the three successive generations, without the selection effect commonly used in a selection breeding program.

Table 1 - Characterization of the places evaluated according to the variables altitude (m), latitude (S) and longitude (W).

<b>Locations</b>	<b>Altitude</b>	<b>Latitude</b>	<b>Longitude</b>
Lavras – MG	918 m	21°14' S	45°00' W
Itutinga – MG	969 m	21°17' S	44°39' W
Ijaci – MG	920 m	21°09' S	44°54' W

Source: By Author (2021)

Table 2 - Description of the experiments (EXP) regarding the crop season (YEAR), inbreeding generations (GENERATION), experimental design (DESIGN), plot size (PLOT) and number of locations (SITES), of replicates (REP), of progenies (PROG) and of checks (CHECK)

<b>EX P</b>	<b>YEAR</b>	<b>GENERATIO N</b>	<b>DESIG N</b>	<b>PLO T</b>	<b>LOCATION S</b>	<b>RE P</b>	<b>PRO G</b>	<b>TES T</b>
A	2016/2017	F <sub>3:4</sub>	Lattice	1m <sup>2</sup>	2	2	194	2
	2017/2018	F <sub>3:5</sub>	Lattice	1.5m <sup>2</sup>	3	3	98	2
	2018/2019	F <sub>3:6</sub>	Lattice	3m <sup>2</sup>	2	3	23	2
B	2016/2017	F <sub>3:4</sub>	Lattice	1m <sup>2</sup>	2	2	194	2
	2017/2018	F <sub>3:5</sub>	Lattice	1.5m <sup>2</sup>	1	3	194	2
	2018/2019	F <sub>3:6</sub>	Lattice	3m <sup>2</sup>	2	3	194	2

Source: By Author (2021)

The climate described for these three sites, according to Martins *et al.* (2018), is classified as tropical altitude, undulating dominant relief and natural vegetation of the “cerrado”. In all environments evaluated, a no-tillage system was used. Sowing was carried out in the first half of November in those environments. Fertilization, inoculation, via furrow, pest, and weed control were carried out in accordance with the recommendation of cultural treatments for soybean cultivation in the region.

The following traits were measured:

- Days to full maturity: number of days from sowing until 90% of the plants of the plot in the R8 stage (MAT) (FEHR; CAVINESS, 1977);
- Yield: grains harvested in the useful area of the plot were weighted and the value converted to kg per hectare after correction to 13% moisture content (YLD).

## 2.2 Statistical analysis of data

### 2.2.1 Phenotypic data analysis

The traits were analyzed via a mixed model approach (BERNARDO, 2020), aiming at the selection/identification the better progenies. For the analysis of the phenotypic data obtained in the two experiments A and B, mentioned above, three different approaches in data analysis were used.

The scenario I was established from Experiment A, a situation in which only the progenies evaluated in each reference generation were evaluated separately. For this analysis the model presented below was used in each generation:

$$y_{ijkl} = \mu + p_i + r_{j(l)} + b_{k(jl)} + a_l + pa_{il} + \varepsilon_{ijkl}$$

where  $y$  is the vector of observed phenotypic values for each trait,  $\mu$  is the intercept of the model,  $p_i$  is the effect of genotype  $i$ ,  $p_i \sim N(0, \sigma_p^2)$ ;  $r_{j(l)}$  is the effect of replicate  $j$  inside environment  $l$ ,  $r_{j(l)} \sim N(0, \sigma_r^2)$ ;  $b_{k(jl)}$  is the effect of block  $k$  inside replicate  $j$  inside environment  $l$ ,  $b_{k(jl)} \sim N(0, \sigma_b^2)$ ;  $a_l$  is the effect of environment  $l$ ,  $a_l \sim N(0, \sigma_a^2)$ ;  $pa_{il}$  is the effect of interaction between genotype  $i$  and environment  $l$ ,  $pa_{il} \sim N(0, \sigma_{pa}^2)$ ; and  $\varepsilon_{ijk}$  is the vector of residuals ( $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ ).

Additionally, data analysis was performed considering experiment B. In this case, full progenies were used during the three agricultural years. This scenario II allowed us to calculate the phenotypic means adjust of all progenies, disregarding the selection effect. The model used in this analysis was the same reported in the scenario I. Excepted in the case of generation F<sub>3:5</sub>, how it was evaluated just one location, the model presented was:

$$y_{ijk} = \mu + p_i + r_j + b_{k(j)} + \varepsilon_{ijk}$$

where  $y$  is the vector of observed phenotypic values for each trait,  $\mu$  is the intercept of the model,  $p_i$  is the effect of genotype  $i$ ,  $p_i \sim N(0, \sigma_p^2)$ ,  $r_j$  is the effect of replicate  $j$ ,  $r_j \sim N(0, \sigma_r^2)$ ,  $r_j$  is the effect of block  $k$  inside replicate  $j$ ,  $b_{k(j)} \sim N(0, \sigma_b^2)$ ; and  $\varepsilon_{ijk}$  is the vector of residuals ( $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ ).

Finally, the third proposed scenario, III, was to obtain the adjusted phenotypic means from a joint analysis of data from progenies was carried out considering the three generations

(F<sub>3:4</sub>, F<sub>3:5</sub>, and F<sub>3:6</sub>), concurrently based in experiments A and B. The model used in this type of strategy is the same described firstly.

In order to evaluate the efficiency of the soybean progeny selection strategy, the genotypes with the highest BLUP (Best linear unbiased prediction) in generation *i* were selected for YLD in three selection intensities (5%, 15%, and 30%) (RAMALHO *et al.*, 2012). The same was done for MAT, but with the intention of reducing the trait, thus selecting the genotypes with lower BLUP. From the phenotypic data, a classification table of selected progenies for each selection intensity in each reference generation was constructed. Those BLUP's were characterized for each scenario proposed according the Table 3.

Table 3 - Description of the BLUP's for each scenario and experiments in the phenotypic data analysis

SCENARIO	EXPERIMENTS		
	A (selection)	B (non-selection)	A and B
I	BLUP-A		
II	BLUP-B		
III	BLUP-J		

Source: By author (2021)

For all data analysis approaches estimates of the genetic and phenotypic parameters, coefficient of experimental variation (*CV*) and selective accuracy ( $r\hat{g}g$ ) were used to assess the experimental quality (RESENDE; DUARTE, 2007).

$$CV = \frac{\sqrt{\sigma_E^2}}{\bar{x}}$$

$$r\hat{g}g = \sqrt{h^2}$$

where:

$\sigma_E^2$ : residual variance;

$\bar{x}$ : mean of variables observed;

$h^2$ : heritability;

The heritability ( $h^2$ ) at the level of progenies were obtained according to the estimator proposed by Lush (1949):

$$h^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{p \times A}^2}{n} + \frac{\sigma_e^2}{nr}}$$

where:

$\sigma_p^2$ : genetic variance;

$\sigma_{p \times A}^2$ : genetics by locations interaction variance;

$\sigma_e^2$ : residual variance;

$n$ : number of locations;

$r$ : number of replicates.

The efficiency of the three scenarios which considered or not the selection effect was evaluated by the realized genetic gain, expected genetic gain, correlated response, and coincidence index based on the BLUP's of the progenies selected under each intensity of selection.

The realized genetic gain ( $RG\%$ ), expected genetic gain ( $EG\%$ ) and correlated response ( $CR\%$ ) were computed for each strategy for subsequent generations according to the following modified expressions of (FALCONER; MACKAY, 1996):

$$RG (\%) = \frac{\overline{BLUP'_{iS_{j/l}}}}{\bar{Y}_j} * 100$$

$$EG (\%) = \frac{\overline{BLUP'_{iS_{j/l}}}}{\bar{Y}_j} * 100$$

$$CR_{y/y'} (\%) = \frac{\overline{BLUP'_{iS_{y/y'}}}}{\bar{Y}} * 100$$

where BLUP's  $j/i$  is the mean of BLUP's of progeny or lines in generation  $j$  considering the selection made in generation  $i$ ;  $\bar{Y}_j$  corresponds to the general mean of progenies or lines in generation  $j$ ; BLUP's  $y/y'$  is the mean of BLUP's of progenies for the  $y$  trait considering the selection made in trait  $y'$  and  $\bar{Y}$  corresponds to the general mean of progenies for trait  $y$ . For the grain yield, progenies were selected to increase the phenotypic expression or higher BLUP's,

while for full maturity, it was desirable to reduce the days to maturity or select progenies with lower BLUP's.

The selection efficiency was verified by the coincidence index proposed by Hamblin and Zimmermann (1986):

$$IC = \frac{(A - C)}{(B - C)}$$

where A: number of progenies matched in the two scenarios; M: number of selected progenies according to the selection proportion; C: number of coincident progeny due to chance. This quantity was calculated by the product of the selection proportion and the number of selected progenies. The IC was calculated considering a two-by-two combination of the three different ways. The first one is inside scenario I, considering two generations, the second one comparing scenario I by II, and the last one comparing the scenario III with the I.

### 2.2.2 Simulation data analysis

The simulation scenario was based on a diploid ( $2n = 40$ ) soybean population, which we obtained using the software PedigreeSIM v2.0 (VOORRIPS; MALIEPAARD, 2012). We designed the genetic linkage map considering the group sizes and centromere distribution based on the Soybean Consensus Map 3.0, produced by Choi *et al.* (2007) and available at the SoyBase platform (GRANT *et al.*, 2009). We generated 100 different biparental populations where parental genotypes were randomly assigned, considering 339 and 325 QTL associated with YLD and MAT, respectively. We distributed QTL along the chromosomes according to their position information (available on SoyBase) and sampled additive effects from an exponential distribution with rates 0.1 and 4 for each trait. For each population, we simulated all model effects considering the following combinations of heritabilities:

Table 4 - Heritability coefficients for YLD and MAT traits used in data simulation.

$h^2$ YLD	$h^2$ MAT
0.3	0.5
0.5	0.5
0.8	0.5
0.3	0.8
0.5	0.8
0.8	0.8

Source: By author (2021)

Given the sampled additive effects, we simulated residual effects from a normal distribution with mean zero and residual variance defined by the targeted heritability level, following the equation:

$$V_e = \left( \frac{V_a}{h^2} \right) - V_a$$

where  $V_e$  is the residual variance,  $V_a$  is the additive variance, and  $h^2$  is the targeted heritability. We then calculated the additive genetic value for each individual by the sum of the products of the simulated genotypes for each loci by their respective additive effects. Then, we obtained the phenotypic means by the sum of the additive and residual effects:

$$y_i = \sum a_i + \varepsilon_i$$

where  $y_i$  is phenotypic mean of the individual  $i$ ,  $\sum a_i$  is the sum of the additive effects for all QTL for individual  $i$ , and  $\varepsilon_i$  is the residual effect for individual  $i$ .

We performed 100 simulations for each heritability combination and base population, totaling 60000 simulations, and constructed each population on top of a pedigree design based on selfings of 200 individuals in the F<sub>1</sub> and F<sub>2</sub> generations, respectively. At generation F<sub>3</sub> we proceeded to individual progeny control, allowing everyone to produce 80 offsprings. We then performed three successive selfing generations using progeny bulks (F<sub>3:4</sub> to F<sub>3:6</sub>) and conducted three selection strategies over the final populations.

We applied seven different selection intensities (1%, 5%, 10%, 15%, 25%, and 30%) at generations F<sub>3:4</sub> and F<sub>3:5</sub>, then compared the rank of selected progenies following these strategies: 1- F<sub>3:6</sub>, in which the progenies were advanced considering the selection effect proposed by experiment A, with the intention of finding the best progenies in F<sub>3:6</sub>, always considering the reference generation; 2 - F<sub>3:6</sub> Full Progenies, a strategy similar to experiment B in which all progenies are selected disregarding the selection effect; and 3 - MGPM, similar to the previous strategy according to experiment A, but when taking data in the last generation (F<sub>3:6</sub>) the information from previous generations is taken into account, simulating a joint analysis. For each strategy, we calculated the associated genetic parameters for all selected progenies, their expected and realized gains (EG; RG). The coincidence indices (CI) was calculated considering the comparisons of strategies 1 vs 2 and 1 vs 3. Correlated responses to



selection for both traits (CR) were performed based in mean of  $F_{3:6}$  progenies of the first strategy.

We performed all analysis using the R v4.0.3 software (R CORE TEAM, 2021) and the packages lme4 (BATES *et al.*, 2015) and lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2017). All graphics were generated with the package ggplot2 (WICKHAM, 2016).

### 3 RESULTS AND DISCUSSION

#### 3.1 Phenotypic data analysis

The summaries of the estimates of genetic and phenotypic parameters from the data analysis were considered in Table 5. The parameters described in the table are presented for the traits YLD and MAT in several inbreeding generations. The results appeared significant variability between progenies considering both traits evaluated, showing the existence of genetic variability between progenies and controls obtained from the biparental cross between P98Y11 x BRS FAVORITA RR. The exception refers to the  $F_{3:5}$  and  $F_{3:6}$  BLUP-A generations for YLD and  $F_{3:6}$  BLUP-A for MAT. Despite the non-significance, differences in means were verified, except in  $F_{3:5}$  BLUP-A, allowing to classify of the best progenies. The applied selection intensities of 50% ( $F_{3:5}$ ) and 25% ( $F_{3:6}$ ) in the experiments classified as "A" may have been drastic enough to exhaust the genetic variance between the progenies, when considering a biparental cross. In addition, when dealing with advanced generations of inbreeding, it is noteworthy that the lack of genetic variability for grain yield may be related to the advancement of generations, since it is expected that  $F_{3:5}$  and  $F_{3:6}$  progenies have a high level of homozygosity, in breeding methods based on hybridization (RAMALHO *et al.*, 2012).

The existing variability observed between progenies can also be corroborated using the frequency distribution graphs (FIGURE 1 and FIGURE 2). It was observed that in the mean amplitudes of all BLUP's associated with the traits, there is genetic variability among the progenies. The phenotype observed in the field is the expression of the genotype, the environment, and the genotype by environment interaction in which the characters of agronomic interest are controlled by hundreds of genes and present a continuous distribution, "quantitative trait" (RAMALHO *et al.*, 2012; SILVA *et al.*, 2017).

In the present work, a biparental cross between elite genotypes (BRS Favorita RR x P98Y11 RR) with excellent agronomic performance in the cultivated region and distinct genetic background was used to obtain the base population. Cultivar BRS Favorita RR is determined growth, maturity group 7.9, and resistance to the root-knot nematode. On the other hand P98Y11 RR presents determined growth, maturity group 8.1, and resistance to the soybean cyst nematode races 1 and 3. The use of elite and contrasting genotypes as parents provided the existence of variability and good agronomic performance among the progenies (BERNARDO, 2020; VAN GINKEL; ORTIZ, 2018).

When considering the interaction between genotypes and environments (G by E), it is possible to observe significance by the likelihood test for both traits. That is, there is a non-coincident behavior between the progenies in the different sites evaluated in each reference generation. Regarding the joint analysis (BLUP-J), the effect of environments is composed of the effect of different locations and different inbreeding generations, and it is not possible to isolate these effects. Obtaining information about the interaction of genotypes by environments is opportune, since the plant breeder isolates the component due to the environment, thus allowing to carry out important tasks such as the selection of superior genotypes, disposal and recommendation due to genetic, and not environmental effects (BIANCHI *et al.*, 2020; EVANGELISTA *et al.*, 2021; GESTEIRA *et al.*, 2018; PEREIRA *et al.*, 2017; SILVA *et al.*, 2017; SOARES *et al.*, 2020).

Accuracy magnitudes were relatively higher for MAT comparing to YLD. When considering the accuracies involving experiment "A", with selection (BLUP-A), it was possible to observe a trend of low magnitudes, as they only considered the generation of reference in the data analysis. And also, as accuracy is directly influenced by heritability estimates, it is suggested that the low values in YLD are due to low magnitudes of genetic variance. Concerning the accuracy values observed, there is a tendency to obtain higher estimates in analyzes involving a greater number of progenies (BLUP-B) especially by increasing the sampling space of the number of progenies evaluated. The highest values of accuracy were detected when adding the effect of agricultural years in the case of joint analysis (BLUP-J).

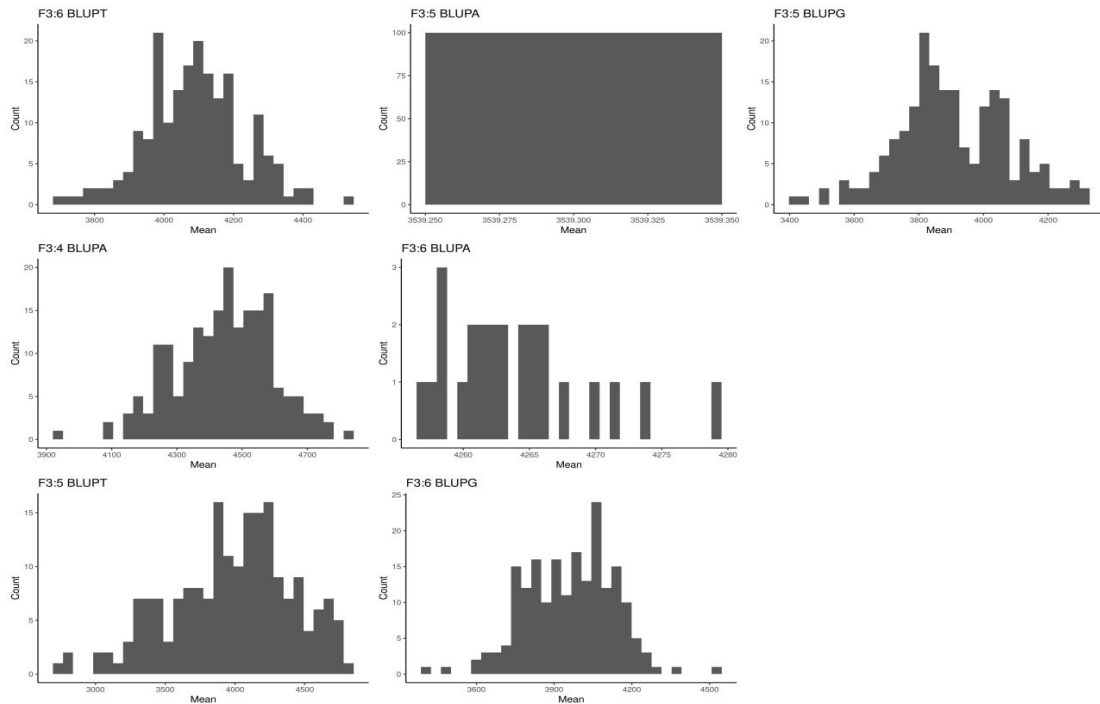
Table 5 - Summary of the joint analysis of genetic and phenotypic parameters associated with agronomic traits grain yield and full maturity among F<sub>3:4</sub>, F<sub>3:5</sub> and F<sub>3:6</sub> soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation.

SOURCE OF VARIATION	GENERATION OF INBREEDING								
	F <sub>3:4</sub> BLUP-A/BLUP-B	F <sub>3:5</sub> BLUP-A	F <sub>3:6</sub> BLUP-A	F <sub>3:5</sub> BLUP-B	F <sub>3:6</sub> BLUP-B	F <sub>3:5</sub> BLUP-J	F <sub>3:6</sub> BLUP-J		
EXPERIMENT EVALUATED	A;B	A	A	B	B	A;B	A;B		
<b>Random effects</b>		<b>YIELD (kg.ha<sup>-1</sup>)</b>							
Progenies (p)	$\hat{\sigma}_p^2$	88781*	0	2154 <sup>ns</sup>	340400***	74649*	74030***	62541***	
Progenies (p) x Environments (a)	$\hat{\sigma}_{p \times a}^2$	178042*	368833***	182037**	-	116620***	168600***	151896***	
Environments (a)	$\hat{\sigma}_a^2$	105423***	578381***	151354*	-	259342***	581400***	434926***	
Residual	$\hat{\sigma}_e^2$	613882	344694	334508	719400	854505	644400	715516	
h <sup>2</sup> (%)		26.79	0	1.44	58.66	42.37	57.53	63.56	
rgg'		0.51	0	0.12	0.76	0.65	0.78	0.80	
CV <sub>e</sub> (%)		17.62	14.74	13.56	23.96	12.45	20.54	21.34	
Mean		4444.64	3982.11	4263.84	3539.31	4089.42	3906.71	3962.53	
<b>Random effects</b>		<b>FULL MATURITY (days)</b>							
Progenies (p)	$\hat{\sigma}_p^2$	36.1562***	9.4921***	2.4714 <sup>ns</sup>	7.7188***	16.2488***	17.282***	17.0928***	
Progenies (p) x Environments (a)	$\hat{\sigma}_{p \times a}^2$	2.2047*	2.9119***	12.7363***	-	3.1635***	1.770*	2.9961***	
Environments (a)	$\hat{\sigma}_a^2$	6.9925.	34.5248***	127.1395***	-	106.1157***	17.939***	32.5888***	
Residual	$\hat{\sigma}_e^2$	11.8229	10.9437	3.0052	6.3714	6.27032***	23.055	16.3775	
h <sup>2</sup> (%)		89.90	81.27	26.46	78.38	86.08	92.39	95.63	
rgg'		0.94	0.90	0.51	0.88	0.92	0.96	0.97	
CV <sub>e</sub> (%)		2.31	2.26	1.24	1.79	1.76	3.29	2.79	
Mean		148.33	146.60	139.71	140.41	142.25	145.57	144.71	

<sup>1</sup>Genetic variance( $\hat{\sigma}_p^2$ ), Genetic variance between progenies and environments( $\hat{\sigma}_{p \times a}^2$ ); Environments variance( $\hat{\sigma}_a^2$ ); Residual variance( $\hat{\sigma}_e^2$ ); Heritability (h<sup>2</sup>); Accuracy selective (rgg'); and Coefficient of variation (CV<sub>e</sub>); Average of BLUPs (Mean)\*Significance by Likelihood Ratio Test - 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.'0.1 '.'1.

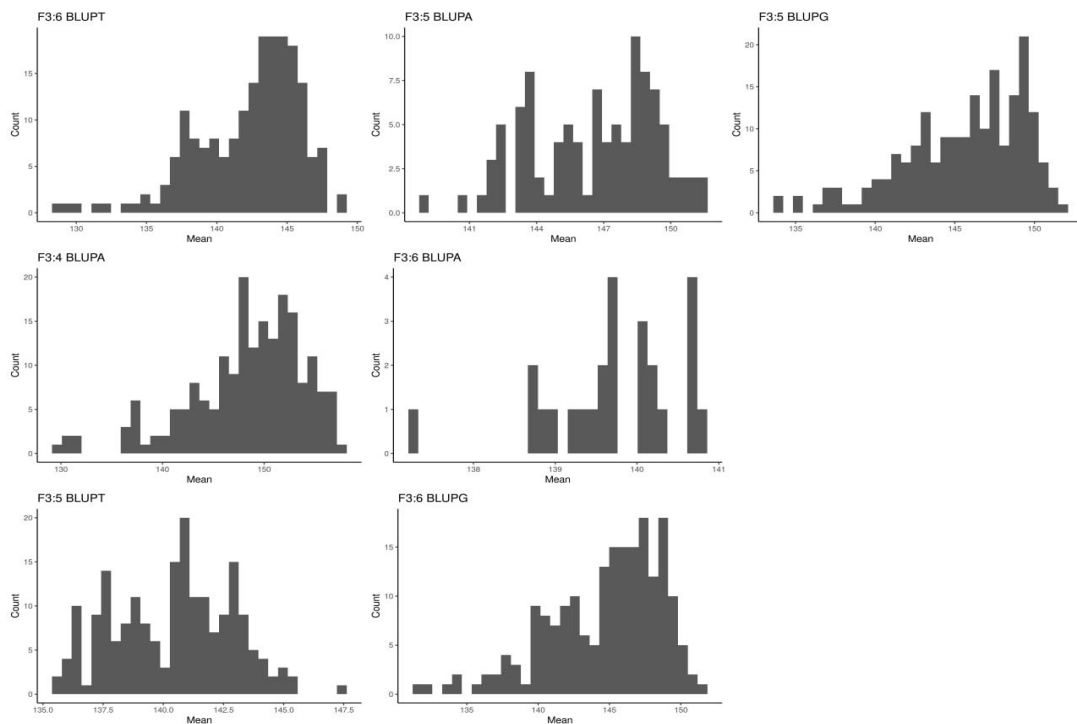
Source: By author (2021)

Figure 1 - Frequency distribution of average BLUPs associated with YLD related  $F_{3:4}$ ,  $F_{3:5}$  and  $F_{3:6}$  soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation.



Source: By Author (2021)

Figure 2 - Frequency distribution of average BLUPs associated with MAT related  $F_{3:4}$ ,  $F_{3:5}$  and  $F_{3:6}$  soybean progenies, considering experiments A and B conducted in different crops season and environments of evaluation



Source: By Author (2021)

On the other hand, the environmental variation coefficient was found values of 1.24% for  $F_{3:6}$  BLUP-A to MAT up to 23.96% in  $F_{3:5}$  BLUP-B to YLD presenting excellent experimental precision MAT, and relative precision for YLD (TABLE 5). As the coefficient of variation is exclusively based on the residual and mean variation of progenies, its magnitudes were considered low, indicating a relative experimental precision. On the other hand, when the experimental precision is measured by estimating the accuracy, the average effect is suppressed, which is one of the main advantages (RESENDE; DUARTE, 2007). These traits are highly influenced by environmental factors and it is noteworthy that similar results were observed when evaluating progenies in the soybean in the same soybean region (BIANCHI *et al.*, 2020; GESTEIRA *et al.*, 2018; RIBEIRO *et al.*, 2020; SOARES *et al.*, 2020).

Heritability magnitudes to YLD ranged from 0 in  $F_{3:5}$  BLUP-A (genetic variance equal to zero) to the value of 63.56% in  $F_{3:6}$  BLUP-J (TABLE 5). When it comes to MAT, the values increase moderately, which may represent greater gains to previous generations in terms of selection of earlier progenies. There heritability magnitudes were ranging from 26.46% in  $F_{3:6}$  BLUP-A, a case of non-significant genetic variance, to 95.63% in  $F_{3:4}$  BLUP-J. The high estimates of heritability for generations that involve a greater number of progenies and/or capture more information in the model, such as the implied year effect, allow more accurate selections to be made within a breeding program because most of the genotype is being represented in the phenotypic information.

For traits of a quantitative nature there are reports in the literature that corroborate the magnitudes found in this work. It was found that the variance components, heritability and coefficient of experimental variation were better estimated when the effect of the population was included, providing greater gain with selection for YLD and MAT (LEITE *et al.*, 2016; PEREIRA *et al.*, 2017; SILVA, 2018).

From the BLUP averages, for each scenario, expected gain, realized gain, correlated response and coincidence index were calculated (TABLE 6). All these estimates were based on generations classified as selection generation and validation generation according to the proposed objective. The first scenario, was based on the experiment called "A", aimed to represent the breeding programs that are based on high selection intensities supported by statistical analyzes considering only the last inbreeding generation, disregarding the year/generation effects. In this scenario, the expected gain with selection was based on the  $F_{3:5}$  BLUP-A generation as selection and validation generation. In other words, it was estimated what is expected to gain with the best progenies in  $F_{3:5}$  in the average of their generation. When

dealing with realized gain, the validation generation becomes the individuals selected in F<sub>3:6</sub> BLUP-A being compared with the average of the progenies in F<sub>3:5</sub> BLUP-A.

It is expected that when increasing the selection intensity, that is, selecting a smaller number of progenies with high performance, the tendency is to obtain greater magnitudes of expected and realized gains. There was no genetic variance for YLD in F<sub>3:5</sub> BLUP-A and with the average of F<sub>3:5</sub> BLUP-A, 3982.11 kg.ha<sup>-1</sup>, the gain realized and expected couldn't be calculated, demonstrating null gains (TABLE 6). For the trait MAT the negative values associated with the estimates of gain show that the objective is to reduce trait expression. The highest magnitudes were identified in selections of 5% when it was selected a smaller number of progenies (TABLE 6). Reports in the literature state that some effects of interaction on the trait under study can under or overestimate parameters such as the expected gain from selection (AMARAL *et al.*, 2019; PEREIRA, L. A.; ABREU; RAMALHO, 2017).

Table 6 - Estimates of the parameters expected gain (EG), realized gain (RG), correlated response (CR) and coincidence index (CI) for YLD and MAT

Trait	Scenario	Parameter	Selection intensity		
			5%	15%	25%
YIELD	I (BLUP-A)	EG (%)	0	0	0
		RG (%)	0	0	0
		CR (%)	20.28	20.39	20.4
		CI (%)	0	0	0
	II (BLUP-B)	EG (%)	19.78	18	16.43
		RG (%)	4.99	4.59	4.26
		CR (%)	5.08	4.81	4.71
		CI (%)	0	6.67	20
	III (BLUP-J)	EG (%)	9.54	7.35	6.5
		RG (%)	10.38	7.24	7.75
		CR (%)	7.82	7.57	6.54
		CI (%)	0	26.66	28
FULL MATURITY	I (BLUP-A)	EG (%)	-3.78	-2.99	-2.59
		RG (%)	-5.71	-5.36	-5.23
		CR (%)	0	0	0
		CI (%)	20	33.33	36
	II (BLUP-B)	EG (%)	-3.24	-2.99	-2.75
		RG (%)	-1.3	-2.12	-2.43
		CR (%)	1.83	1.14	1.38
		CI (%)	0	6.67	20
	III (BLUP-J)	EG (%)	-7.96	-6.11	-5.09
		RG (%)	-9.94	-7.63	-6.96
		CR (%)	-3.97	-2.75	-2.6
		CI (%)	20	6.66	16

Source: By author, 2021.

Still concerning first scenario, it was possible to perform the correlated response by selecting the progenies with the highest YLD and the response for the MAT, and the opposite.

When selecting for YLD in F<sub>3:5</sub> BLUP-A, the response in MAT in the F<sub>3:6</sub> BLUP-A generation did not vary with different selection intensities because of the non-variance in YLD. On the contrary, in the selection involving MAT in F<sub>3:5</sub> BLUP-A it was possible to find a response positive in the YLD in the generation F<sub>3:6</sub> BLUP-A (TABLE 5). The parents used to obtain this population are highly productive and long full maturity, which justifies the low variations in the correlated response.

According to Bernardo (2020), from the positive association and high magnitude, it is possible to select a trait that is difficult to measure or has low heritability, through an easily measured trait. There is evidence that the selection of progenies with late full maturity and longer reproductive period results in higher yields (GESTEIRA, 2018; SOARES *et al.*, 2020). Considering that in Brazil it is possible to carry out more than one crop per agricultural year, the adoption of early and also productive soybean cultivars would enable the sowing of the successor crop in the production system to be carried out earlier under more favorable conditions, that is, better water availability, average temperatures more suitable for plant development and, in consonance, a better maximization of the production system (GESTEIRA, 2015).

Finally, still considering the first scenario, the coincidence index was determined based on the ranking of the selected progenies in F<sub>3:5</sub> BLUP-A and the ranking in the validation generation F<sub>3:6</sub> BLUP-A. Then, as the selection intensities progress, the index tends to increase when more progenies are selected (TABLE 6).

In relation to the second scenario proposed in the methodology, the results are based on the analyzes that permeate experiment “B”, that is, the whole 196 progenies from the taking away of plants from population F<sub>3</sub> were carried to advanced generations of inbreeding. It should be noted that this option for conducting by breeders is not idealized but was carried out to obtain phenotypic data from the unselected progenies.

About gains, the selection generation was F<sub>3:5</sub> BLUP-B and the progeny validation generation refers to F<sub>3:6</sub> BLUP-B. For YLD, gains increased when selecting based on all progenies, proving that the inclusion of unselected progenies increases the selection sampling space, generating greater genetic selection variability resulting in greater gains. In the case of correlated responses, it is observed that when evaluating all progenies, the most productive progenies generated a positive response in the number of days to full maturity (TABLE 6).

Differently from the previous parameters, to calculate the coincidence index, the F<sub>3:5</sub> BLUP-B generation was used as the selection generation and its corresponding coincidence was

estimated by the  $F_{3:6}$  BLUP-A, which generation considers the selection effect. This alignment was defined to compare the two-generational advancement strategies in the breeding program. It is noticed that in this scenario only MAT was used with the comparison intuit with the previous scenario since for YLD the coincidence values were null. The coincidence drastically decreases, indicating a change in ranking and that the progenies that were no longer selected in the first generations of inbreeding, could have been selected in future generations from the phenotypic data obtained by the experiments categorized as B (TABLE 6).

In the third scenario, the year effect was considered in the analyzes since the BLUP-J estimates were made from a joint analysis of all generations. The database includes experiments A and B, through the mixed model approach due to data imbalance caused by different generations proposed in this scenario. Thus, the selection generation of this scenario was the  $F_{3:5}$  BLUP-J which obtains all phenotypic data up to the  $F_{3:5}$  generation in the two experiments and the validation generation was the  $F_{3:6}$  BLUP-J, like the previous one.

Overall, the realized gains were greatest for YLD and MAT in all selection intensity. By including the year effect in the evaluation, it is possible to obtain better estimates by not including only the reference generations as selection. Thus, it is evident that the inclusion of the year effect allows, in addition to improving parameter estimates, to obtain greater gains with selection. Similar study considering the population effect was conducted by Pereira *et al.* (2017). In this case, coincidence and ranking among the selected progeny with and without the effect of the population are of greater magnitude in more advanced inbreeding generations and at higher percentage of selected progeny.

For the coincidence index parameter, the  $F_{3:5}$  BLUP-J (selection generation) was compared with the  $F_{3:6}$  BLUP-A (validation generation). The results of this scenario was slightly larger with the previously, indicating a better coincidence with the evaluations that were based on reference generations of experiment A. As we used both experiments A and B to calculate the BLUP's estimates of the joint analysis, a greater traceability (repeated) of database from experiment A in the selection of the best ones in this type of analysis may have led to this result. However, it shows, together with the estimated phenotypic and genetic parameters, that the ranking and for future selection of the best progenies becomes interesting to use different agricultural year.

Table 7 allows a brief analysis of the results of the phenotypic data. In summary, the analyzes that do not only consider the effects of selected progenies by each reference generation commonly represented in a traditional plant breeding program, had parameter estimates,



coincidence indices, and genetic gains different from the proposed alternative methodologies. There are reports that large soybean breeding programs test 200 F<sub>2</sub> populations, advance in bulk up to F<sub>4</sub> and tests an average of 350 progenies/cross (BERNARDO, 2014). This drastic sampling may be having a severe impact on obtaining superior progenies as they often consider only a reference generation and do not consider the information from the unselected progeny. It is known that the selected genotypes have a high number of information associated with them, which allows the adoption of information from all generations to select the best lines, although this strategy has not been widely used by breeding programs.

With the inclusion of unselected progenies, it was found that the precision estimates improve when comparing the parameters of accuracy, heritability and coefficient of variation in the respective inbreeding generations classified as BLUP-A (with selection) and BLUP-B (no selection) for both traits (TABLE 7). The wide aspect of progenies evaluated from the same biparental crossing contributed to increase the genetic variability of selection. Furthermore, as seen by the reduction in coincidence indices when considering scenarios, I and II, it is suggested that changing the ranking by non-coincidence allows for estimating greater gains, especially about MAT in this case.

In the literature, Resende *et al.* (2015) proposed the use of a multigeneration index, which considers the whole information across generations and the covariance among successive generations with a view toward the identification of the best progenies. The results of the simulation and analysis of field data from *Phaseolus vulgaris* L. were concordant and showed that the estimates of the accuracies were near 85% using the multigeneration index and 70% for the traditional methodology based on comparisons among progeny means of the generation F<sub>3,6</sub> or the last generation evaluated. Another advantage of the index is mitigation of the effect of the genotypes by environments interaction, because it considers all the generations of assessment, thus involving the interaction of the progenies with the sowing times. When selection is performed using only the information from the last generation assessed, the performance of the progenies is inflated by this interaction. The efficiency of the strategy proposed leads to a 12% gain concerning that which is normally used. It should be highlighted that this increase is obtained without any additional cost on experimentation, but simply through changing the procedure of identification of the best progenies after successive generations.

Table 7 - Summaries of the estimates of the parameters heridability ( $h^2$ ), accuracy (rgg'), coefficient of variation ( $CV_e$ ) for the different inbreeding generations advance strategies and coincidence index (CI) and realized gain (GR) at the levels of intensities of selection for the three possible scenarios considering YLD and MAT.

TRAITS	PARAMETERS	SCENARIOS	F <sub>3:4</sub> BLUP-A/BLUP-B	F <sub>3:5</sub> BLUP-A	F <sub>3:6</sub> BLUP-A	F <sub>3:5</sub> BLUP-B	F <sub>3:6</sub> BLUP-B	F <sub>3:5</sub> BLUP-J	F <sub>3:6</sub> BLUP-J
	$h^2$ (%)		26.79	0	1.44	58.66	42.37	57.53	63.56
	rgg'		0.51	0	0.12	0.76	0.65	0.78	0.80
	$CV_e$ (%)		17.62	14.74	13.56	23.96	12.45	20.54	21.34
			<b>Selection intensity</b>						
				<b>5%</b>		<b>15%</b>		<b>25%</b>	
<b>YIELD</b>	<b>IC</b>	<b>I</b>		0		0		0	
		<b>II</b>		0		6.67		20	
		<b>III</b>		0		26.66		28	
	<b>GR</b>	<b>I</b>		0		0		0	
		<b>II</b>		4.99		4.59		4.26	
		<b>III</b>		10.38		7.24		7.75	
	<b>h<sup>2</sup></b>			81.27		78.38		92.39	
	<b>rgg'</b>			0.90		0.88		0.96	
	<b>CV</b>			2.26		1.79		3.29	
			<b>Selection intensity</b>						
				<b>5%</b>		<b>15%</b>		<b>25%</b>	
<b>FULL MATURITY</b>	<b>IC</b>	<b>I</b>		20.00		33.33		36.00	
		<b>II</b>		0.00		6.67		20.00	
		<b>III</b>		20		6.66		16	
	<b>GR</b>	<b>I</b>		-5.71		-5.36		-5.23	
		<b>II</b>		-1.30		-2.12		-2.43	
		<b>III</b>		-9.94		-7.63		-6.96	

Source: By author, 2021.

Also considering Tables 7 of the summarized results, the improvement in data analysis is identified from the highest magnitudes of accuracy and heritability when comparing  $F_{3:6}$  BLUP-A and  $F_{3:6}$  BLUP-J. Another parameter that allows us to verify the advantages of the analysis that considers all previous information about the progenies is related to the greater magnitudes of realized gains.

In the context of a joint analysis of the data, the influence of environmental factors, that is, locations and agricultural years in the phenotypic manifestation is expected. Environmental factors can be classified as predictable and unpredictable. The predictable act systematically as soil type and actions determined by man such as planting date, plant stand, harvesting, among others. The unpredictable factors are those that fluctuate according to the agricultural year and places such as rainfall, hail, frost, temperature, and others (ALLARD; BRADSHAW, 1964; BAROT *et al.*, 2017; KIÆR; SKOVGAARD; ØSTERGÅRD, 2009; MARSHALL; BROWN, 1973; VAN EEUWIJK; BUSTOS-KORTS; MALOSETTI, 2016).

When experimental work involving agricultural generations/years is carried out, preferably in the same location, they usually allow inferences about the difficulty of handling the effect of environments, especially for the source of variation: progenies by agricultural years that have high magnitudes. This fact was frequently reported in experiments with the common bean crop conducted in the same region where the data with the soybean crop from this study, also autogamous, were obtained (FERREIRA *et al.*, 2015; LEMOS; ABREU; RAMALHO, 2020; LIMA; RAMALHO; ABREU, 2012; PEREIRA, L. A.; ABREU; RAMALHO, 2017; RAO *et al.*, 2002).

Certain proposals have emerged aimed at improving the selection process through analyses, primarily using mixed models, drawing on information from all the progenies evaluated in previous generations. Authors include the use of pedigree information to model and exploit genetic correlation among relatives and the use of flexible variance-covariance structures for genotype-by-environment interaction (PIEPHO *et al.*, 2008). The use of previous information from the breeding program, in the case of BLUP-J, is only possible via the mixed model approach due to the unbalance of data caused by the advance of generations. Mixed models have facilities for modeling heterogeneity of genetic variances and correlations between environments as well as for modeling design features and spatial trends in individual trials. Furthermore, prediction of phenotypic traits across environments and estimation of quantitative genetic parameters as genetic variance and correlation, heritability, and responses to direct, indirect and index selection are natural within the context of mixed models, whereas they

become contrived in the context of models with only fixed terms (PIEPHO *et al.*, 2012; VAN EEUWIJK; BUSTOS-KORTS; MALOSETTI, 2016).

Recently published studies show that if an evaluation were to consider all relevant generations, the gain from selection could be higher, especially when an index that involves information from the population that gave rise to the progenies is used. Some authors aimed with this study to compare three procedures in the evaluation of successive generations and to discuss the implications of the progeny by environment interaction in terms of the success of selection. They found that the classification of the progenies by the BLUP's and progenies within populations effect did not expressively differ from that obtained when using only the mean, even when several generations were considered in the selection. Also, none of the procedures used effectively mitigated the effect of the progeny by environment interaction (LEMOS; ABREU; RAMALHO, 2020).

### **3.2 Computer simulation data analysis**

Considering that the data simulation process does not perfectly represent a real phenotypic situation, some strategies were adopted to look up the least possible bias. By using QTL associated with grain yield and full maturity data previously defined enclosed the soybean genome for the traits according to the “Soybase Data”, it allowed us to approach reality about possible recombination of genes. This is because it was possible to represent the distances of genes regarding the centromeres of chromosomes and chiasm frequencies according to the distances between the genes themselves. It is important to note that two genes less than 50 cM far away on the same chromosome are linked and will hardly recombine, remaining in linkage disequilibrium (LD).

The linkage disequilibrium is used to emphasize that gene linkage induces the population to waste more generations to achieve equilibrium because it occurs in independent genes. Thus, the linkage disequilibrium measure is influenced by the recombination frequency and by the information about when and wherein the genealogy of the combinations the mutations occurred (HARTL; CLARK, 2010). Recombination is the main phenomenon that reduces intrachromosomal LD. In addition, population size is important as LD can be created in populations that have undergone a drastic reduction in size due to genetic bottleneck as selection causes genetic narrowing to specific loci and can create LD between the selected allele and the locus switched on. Thus, selection for or against a phenotype controlled by two unlinked genes (epistasis) can create LD although they are not physically linked (HEDRICK, 2009).

In the simulation, the estimation of the individual phenotypic values was performed by random sampling of an exponential distribution of the additive effects of the genes related to the genetic parameters. The exponentiality occurred because it is two quantitative traits that have small effects in a large amount. From the randomization mentioned above, a bias was added that allowed the estimation of the environmental variance weighted by the heritability of the effect. As a result of this sum, everyone had an estimated phenotypic value that together gave a normal distribution for YLD and MAT. The average of the phenotypic values of everyone in their respective generation resulted in the average of the progeny used to calculate gains and correlated responses. As soybean is an autogamous plant, we can consider the average degree of dominance equal to zero. In this situation, the genetic variance will depend only on  $V_A$  (BERNARDO, 2020).

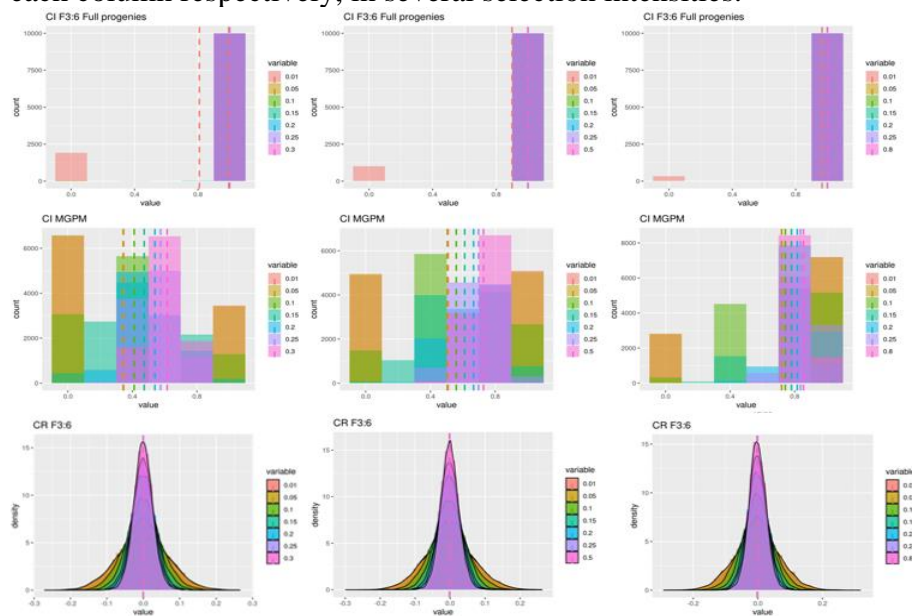
In the analysis of simulated data, strategy 1 ( $F_{3:6}$ ) referred to a lighter selection of 50% of the progenies in the first generation and a more intense selection of 25% of the progenies in the second generation (Experiment A). In a breeding program, the breeder must make decisions especially related to the conduct of progenies from the selection of plants within the bulk, a methodology proposed by Frey (1954). The increase in homozygosity combined with the advancement of inbreeding generations allows to improve the precision in the identification of superior progenies, obtaining one or more inbred lines, from each progeny, to participate in the yield experiments in final phases of the breeding program. Many breeders use this strategy to advance the progenies from phenotypic evaluation experiments that allow selections during inbreeding phases, unlike the original proposal, allowing to optimizing resources such as area availability, for example.

In strategy 2 (Full Progenies), the simulation allowed to proceed with the full progenies obtained from the segregating populations since the beginning (Experiment B). This simulation allows not only to advance with progenies that would be selected in the first generations but also to advance with the possible unselected genotypes. Private breeding companies are known to make drastic selections from one reference generation to another in a breeding program. Cooper (1988) in the soybean breeding program, mentions that he performed 100 crossings/year and from each crossing, he used 300 plants in the  $F_2$  generation. However, it advanced 30 progenies/cross with 1 repetition, a useful plot of 2.5m, that is, approximately 750 plants per population in the  $F_3$  generation. In principle, this last sampling is more drastic than the  $F_2$  generation. More recently, there are reports that large soybean breeding programs evaluate 200

F<sub>2</sub> populations, advance in bulk up to F<sub>4</sub>, and evaluates an average of 350 progenies/cross (BERNARDO, 2020).

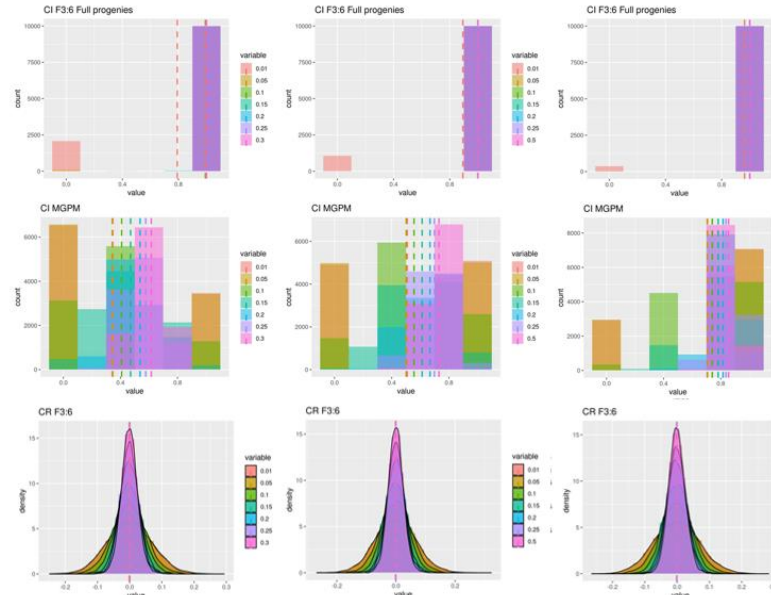
Between the two strategies mentioned above, the computer simulation allows us the possibility to identify high coincidences for the two traits. These coincidences are represented by the graph "CI F<sub>3:6</sub> Full Progenies" for each trait, and for the several heritability combinations. Each column of graphs in each figure represents the tendency of higher accuracy estimated from left to right (FIGURE 3, 4, 5, and 6). When there is a low selection index, that is, the selection is more rigorous and fewer progenies are selected, a smaller coincidence between the two scenarios was observed. It was possible to conclude that the ranking of the best F<sub>3:6</sub> progenies that were selected during different generations diversifiers in relation to the ranks of the best F<sub>3:6</sub> progenies that were not selected during the conduct of the breeding program. However, as this shift was only for strict selection indices, it is expected that the most promising progenies will always be selected in both situations.

Figure 3 - Cluster of graphs of simulated YLD data about coincidence indices between strategy 1 and strategy 2 (CI F<sub>3:6</sub> Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in F<sub>3:6</sub> (CR F<sub>3:6</sub>), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



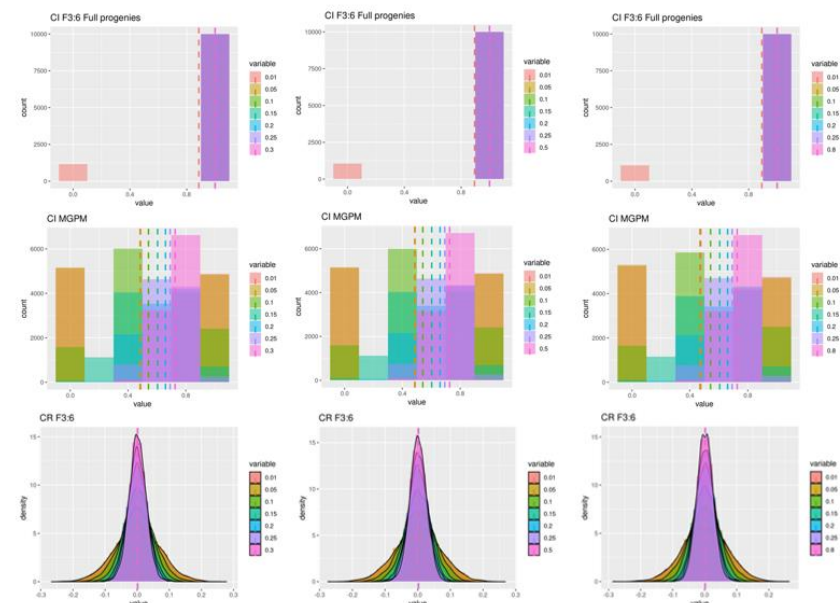
Source: By Author (2021)

Figure 4 - Cluster of graphs of simulated YLD data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities



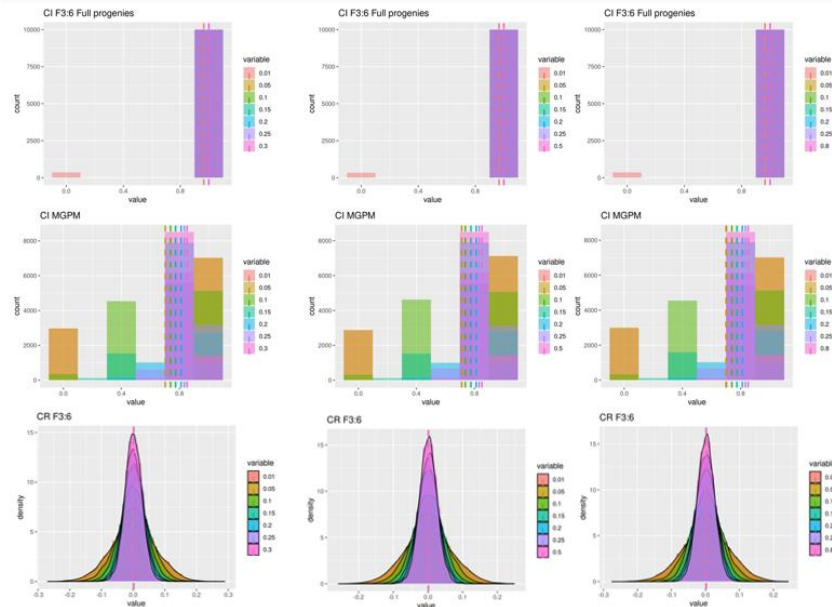
Source: By Author (2021)

Figure 5 - Cluster of graphs of simulated MAT data about coincidence indices between strategy 1 and strategy 2 (CI  $F_{3:6}$  Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in  $F_{3:6}$  (CR  $F_{3:6}$ ), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



Source: By Author (2021)

Figure 6 - Cluster of graphs of simulated MAT data about coincidence indices between strategy 1 and strategy 2 (CI F<sub>3:6</sub> Full Progenies), coincidence indices between strategy 1 and strategy 3 (CI MGPM), correlated response in F<sub>3:6</sub> (CR F<sub>3:6</sub>), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



Source: By Author (2021)

However, the non-coincidence effect explained above is more visible in combinations of low heritability. Taking as an example in which the heritability for YLD was 0.3 and MAT 0.5, almost 25% of a sample space of 10,000 simulations (100 crossings in 100 replicates) had non-coincident progenies at high selection intensities (FIGURES 3, 4, 5, and 6). This situation occurred because the bias of a random characteristic added to the individual's phenotypic value is greater due to the character of low heritability with greater magnitudes of residual variance. Otherwise, in high heritability, the coincidence rate is much higher. Heritability estimates the reliability of phenotypic value as an indicator of reproductive breeding value and is directly related to the accuracy, being one of the most useful genetic parameters for breeders (RAMALHO *et al.*, 2012).

It is important to note that although the two strategies have coincidence progenies, F<sub>3:6</sub> progenies that were not selected in early generations of the breeding may be better than progenies that have gone through two generations of selection. Thus, especially for quantitative traits in which the evaluation accuracy is restricted because they have great environmental influences promising progenies may not be selected in early generations, especially in case of intense selections.



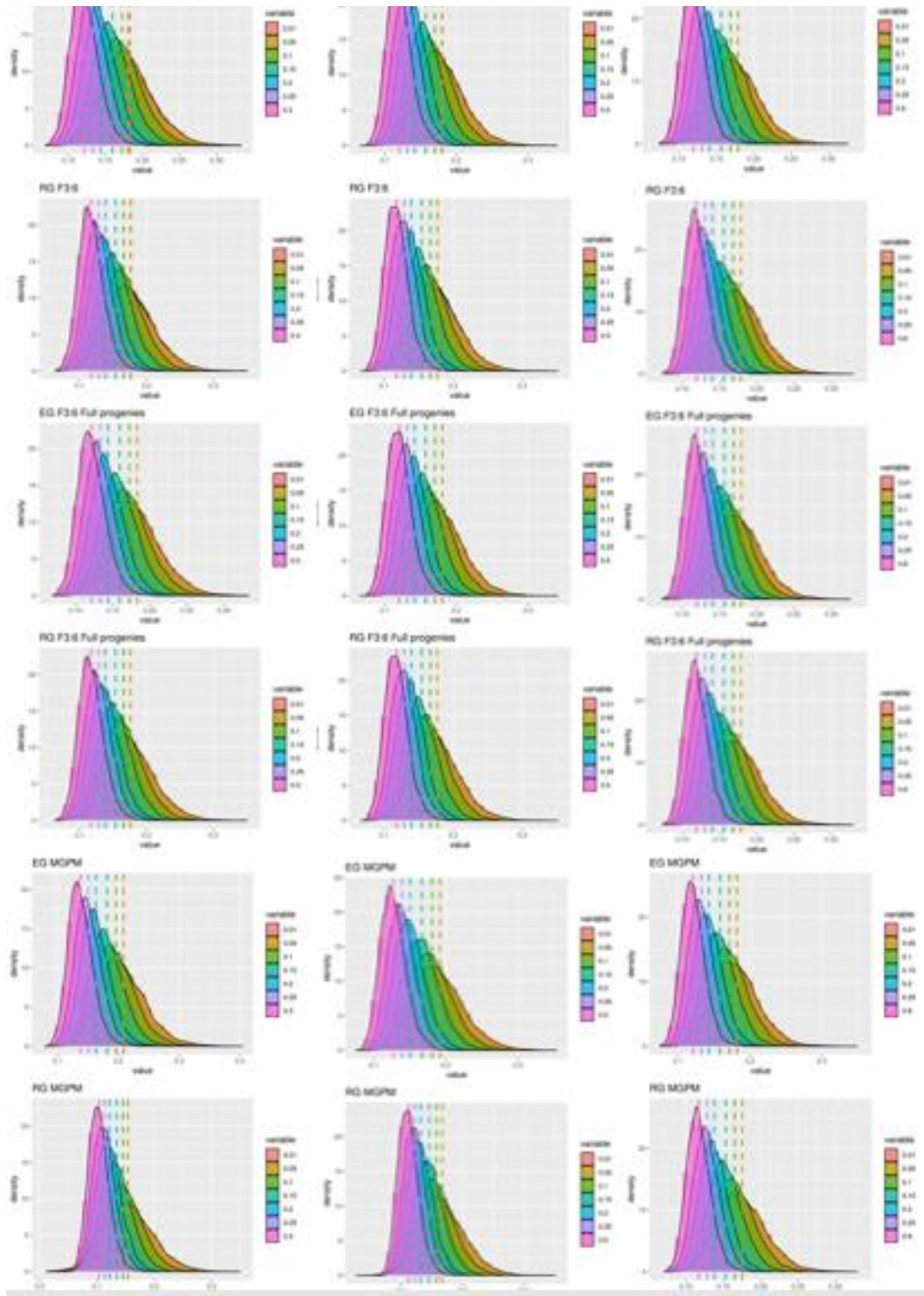
Another situation verified with the data simulation was when it emulated the means of the  $F_{3:6}$  progenies, in a conduction scheme similar to experiment A, that is, with selection, but considering the phenotypic values of all generations ( $F_{3:4}$ ,  $F_{3:5}$ , and  $F_{3:6}$ ). This strategy 3 was denominated MGPM (Multiple Generation Progenies Mean). In this case, the “CI MGPM”, the second line of the cluster of graphs, showed the coincidence indices of this selection strategy with strategy 1, which differently estimated the mean only based on the values of the progenies in the  $F_{3:6}$  reference generation (FIGURES 3, 4, 5, and 6). For the YLD, there was a discrete distribution of the different scenarios in the space of intensities from 0 to 1. This non-regularity of the data was more visible in the low heritability combinations. It is possible to infer that there was a very low coincidence between the selection of the best  $F_{3:6}$  progenies when considering only the information from one agricultural year compared to the average of the progenies considering a greater number of generations.

In plant genetic breeding, progenies will be advanced only by generating reference or at some situation using information from the previous year, such as in Cultivation and Use Value (VCU) experiments. To assist in the selection of the best progenies and/or genotypes, new tools were launched by science, including the application of the mixed model approach (HENDERSON, 1984; MRODE; THOMPSON, 2005). One reason for the use of this methodology in the data analysis of breeding programs is due to the fact that it allows the handling of unbalanced data since the loss of plots and/or progenies is not uncommon to occur under field conditions and also the advance of progenies during successive agricultural years can generate different evaluation designs. Other advantages are allowing the estimators of additive breeding values (VGA) or “breeding values” of progenies (BLUB - best linear unbiased predictor) to be predicted with high accuracy and the incorporation of kinship information in the analyzes (BERNARDO, 2010; BRUZI, 2008; LEMOS; ABREU; RAMALHO, 2020; PIEPHO *et al.*, 2012; RESENDE *et al.*, 2015).

When comparing strategies 1 and 3, it is known that a bias due to generations may have led to a discrepancy in the rankings of the progenies (FIGURES 3, 4, 5, and 6). This is because the situations are distinguished in the genotypes that were used to obtain the means of the progenies. In early generations such as  $F_{3:4}$ , the genetic recombination rate is still lower than those found in advanced generations, culminating in a different variance than that found in  $F_{3:6}$  for example. This recombination allows breaking linking blocks promoting an increase in genetic variance and relative frequency of genotypes containing favorable alleles, as mentioned earlier in this chapter of results.

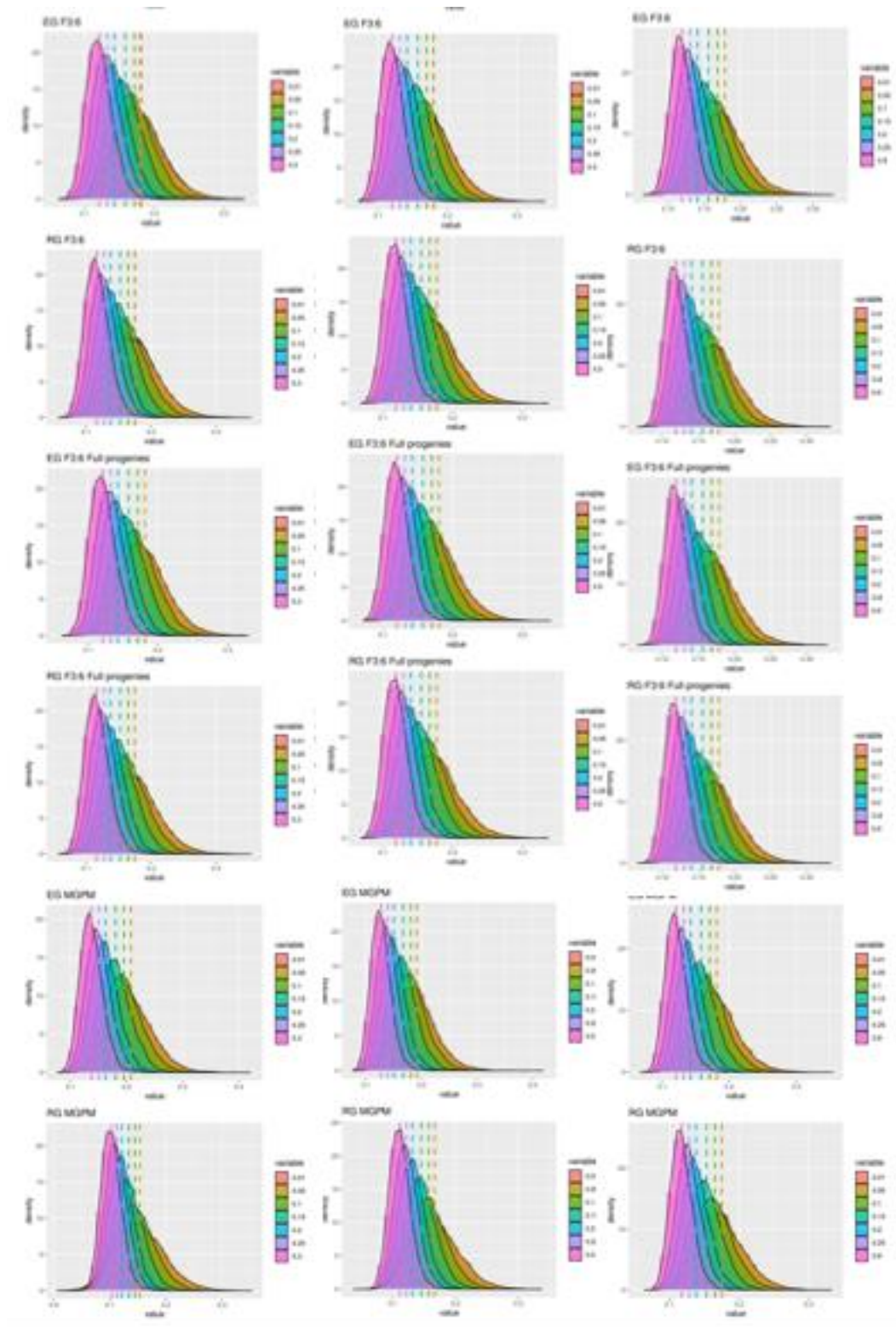
Another aspect to be considered in these two strategies (1 and 3) is related to the fact of the composition of the year effect. The use of three generations generates an important interaction bias, which is the variable year directly related to genotype by environment interaction. In the soybean crop, some work was carried out to study adaptability, stability, and genotypes by environments interaction, showing that the behavior of the genotypes was not coincident in the different environments evaluated, especially due to the different agricultural years of evaluation (BIANCHI *et al.*, 2020; GESTEIRA *et al.*, 2018; RAMALHO; MARQUES; LEMOS, 2021; SOARES *et al.*, 2020; VAN EEUWIJK; BUSTOS-KORTS; MALOSETTI, 2016). Thus, the non-coincidence between the two situations may be related to different averages due to the generations that were evaluated and, therefore, the increase in the bias of agricultural years.

Figure 7- Cluster of graphs of simulated MAT data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



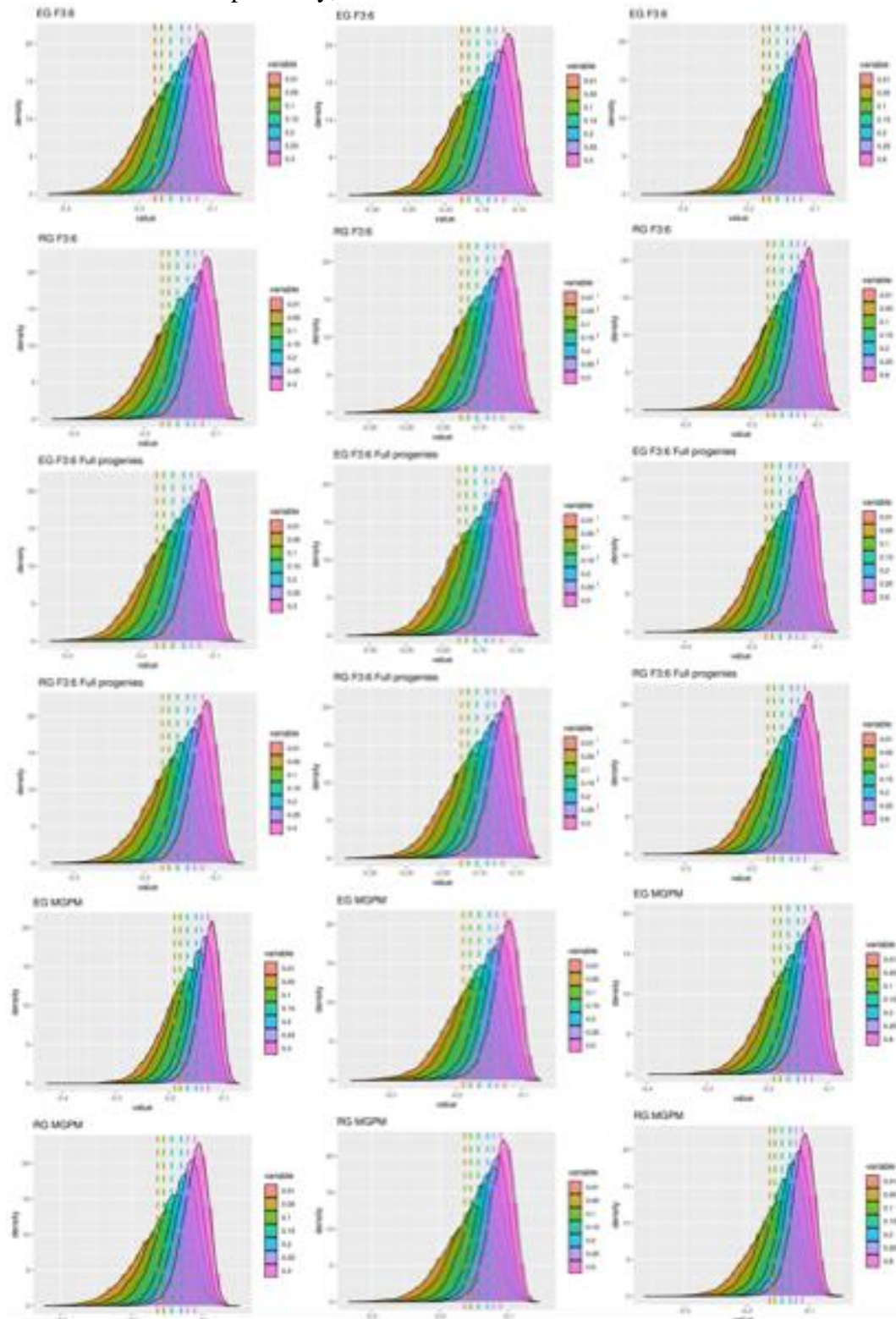
Source: By Author (2021)

Figure 8 - Cluster of graphs of simulated MAT data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



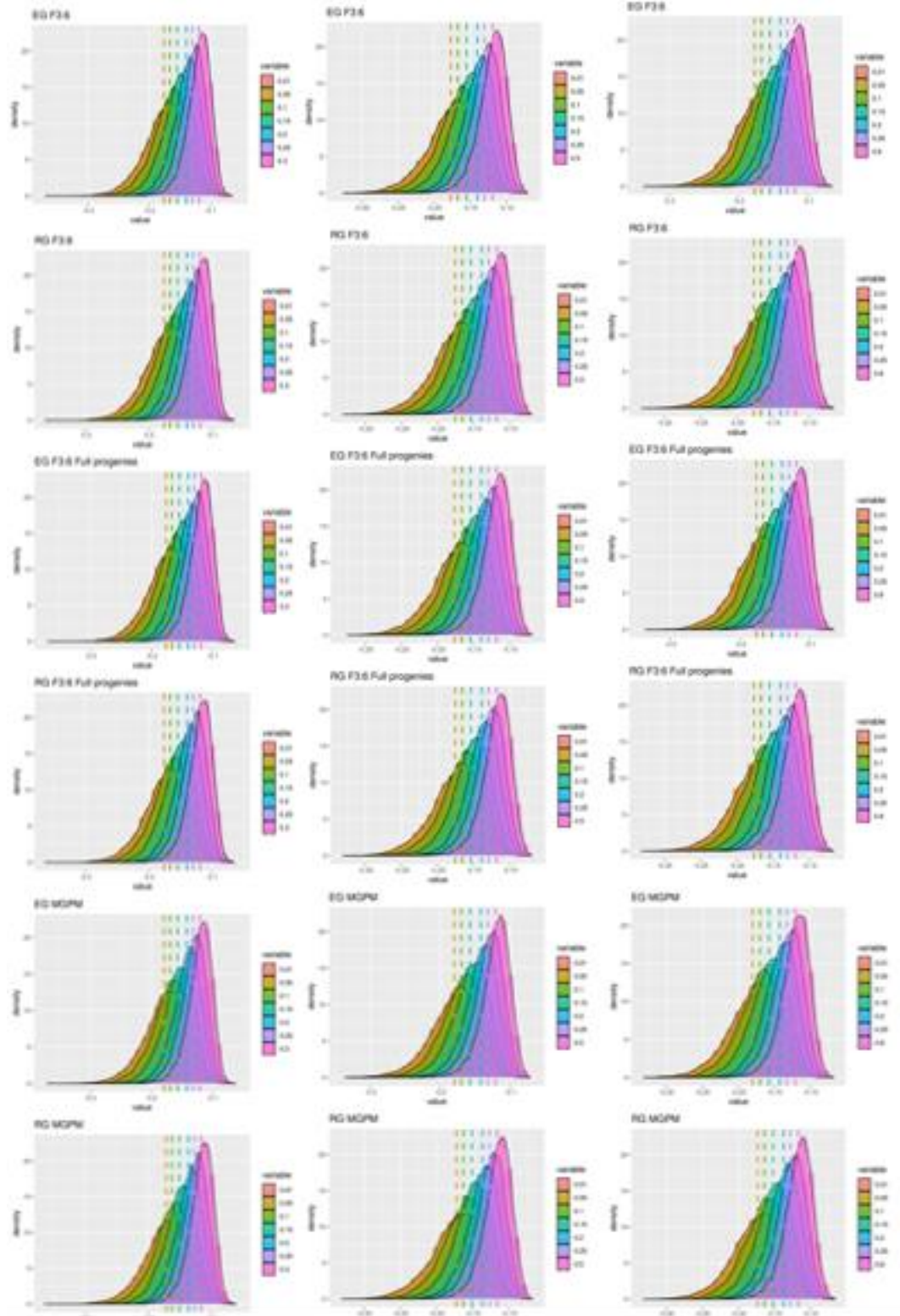
Source: By Author (2021)

Figure 9 - Cluster of graphs of simulated YLD data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.5 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



Source: By Author (2021)

Figure 10 - Cluster of graphs of simulated YLD data about gain realized and gain expected in strategy 1 ( $F_{3:6}$ ), strategy 2 ( $F_{3:6}$  Full Progenies), and strategy 3 (MGPM), considering the heritability of 0.8 for MAT and variation of YLD 0.3, 0.5, and 0.8 represented in each column respectively, in several selection intensities.



Source: By Author (2021)

These gains realized and expected are represented in the graphs of realized gain (RG) and expected gain (EG) for each selection strategy used:  $F_{3:6}$ ;  $F_{3:6}$  Full Progenies; and MGPM (FIGURE 7, 8, 9, and 10). A normal distribution in all situations was identified as expected by the quantitative nature of the traits. When considering the high selection intensities that consider a high number of selected progenies, a decrease in the magnitude of gains can be recognized. Analogously, the smaller number of selections of the best progenies returns superior gains with the advancement of generations.

The realized gain was calculated based on the  $F_{3:6}$  generation in relation to the mean of the  $F_3$  population estimated through the simulation. As for the expected gain, the mean of the progenies in  $F_{3:5}$  was considered, also in concordance with the mean of the  $F_3$  population. These normal distributions make perfect sense because they allow us to visualize the variation in the position of the normal distribution and the direction of the tail according to the evaluated character. That is, for YLD as the gains are positive, the distribution will be positively asymmetric. As for the MAT, a negative asymmetric distribution was found, since the gains are responsible for reducing the number of days to full maturity.

In any selection strategy, the attribute of success comprises the quantification of the expected gain with the selection. Higher selection intensities, that is, when fewer individuals were selected, a greater expected gain was obtained, since individuals with the highest frequency of favorable alleles are selected. However, when using a greater intensity of selection, there is a reduction in variability (AMARAL *et al.*, 2019).

In relation to the correlated responses (CR Graphs), a symmetrical continuous distribution around the zero position was observed (FIGURE 3, 4, 5, and 6). This implies that for most situations, in the process of simulating data the mean of the progenies there was no correlation between the two traits. With that, it was not verified a possible elongation of the cycle verified in genotypes with high yield. Recently, Todeschini *et al.* (2019), evaluating soybean cultivars cultivated in several decades for southern Brazil, describes that there is a reduction in the vegetative period (PV) at the expense of the reproductive period (PR) due to several selection cycles in which the PR/PV ratio increased, corroborating the increase in grain yield. Note that higher yields are associated with late progenies with longer reproductive periods. However, it is possible to select early progenies above average for both traits.

The estimate of the correlated response was performed based on the average of the progenies in  $F_{3:6}$ . It is known that with the advance of generations during the simulation process, the QTL were recombined, breaking the linked blocks that ensured the negative correlation

between yield and full maturity. In case of QTL, its effects are generally estimated in initial generations of recombination (F<sub>2</sub> or BCK), as these guarantee larger linked blocks due to lesser recombination, making possible a better identification of them.

#### **4 CONCLUSIONS**

From the interface between the results involving the phenotypic field data of the evaluated progenies in three generations by different scenarios presented and the results obtained with the estimate of predicted means of the genotypes via computer simulation of data, it is possible to validate some conclusions of this work.

A first approach believes that the use of selected and unselected genotypes in a soybean breeding program with the progress of generations allows drastically to change the ranking, generating a bias in the selection of the best progenies with the advancement of inbreeding generations. This behavior was observed both in the field data evaluations and in the data simulation aspect (cross validation).

Evaluating progenies in order to represent the conditions under which they will be used in the future, thus mitigating the effects of the genotype x environment interaction is extremely challenging on the part of breeders, highlighting the contribution of influencing the interaction per crop season recommended by unpredictable factors.

As a suggestion in decision making for breeders in breeding programs that evaluate thousands of progenies in different locations and, especially agricultural years, it is suggested that when performing statistical analysis of the data, previous information of the progenies should be included via approaches that consider the imbalance of data, given that this alternative does not entail additional costs for the improvement program.



## REFERENCES

- ABREU, G. B. *et al.* Strategies to improve mass selection in maize. **Maydica**, Bergamo, v. 55, n. 3–4, p. 219–225, 2010.
- ALLARD, R. W.; BRADSHAW, A. D. Implications of genotype-environmental interactions in applied plant breeding. **Crop Science**, Madison, v. 4, n. 5, p. 503–508, 1964.
- AMARAL, L. de O. *et al.* Pure line selection in a heterogeneous soybean cultivar. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 19, n. 3, p. 277–284, 2019.
- BAROT, S. *et al.* Designing mixtures of varieties for multifunctional agriculture with the help of ecology. A review. **Agronomy for Sustainable Development**, Paris, v. 37, n. 2, p. 1–20, 2017.
- BATES, D. *et al.* Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, [s. l.], v. 67, n. 1, 2015.
- BERNARDO, R. **Breeding for quantitative traits in plants**. 2. ed. Woodbury: Stemma Press, 2010.
- BERNARDO, R. **Breeding for quantitative traits in plants**. 3. ed. Woodbury: Stemma Press, 2020.
- BERNARDO, R. **Essentials of plant breeding**. Woodbury: Stemma Press, 2014.
- BIANCHI, M. C. *et al.* Heritability and the genotype  $\times$  environment interaction in soybean. **AgroSystems, Geosciences & Environment**, [s. l.], v. 3, n. 1, p. 1–10, 2020.
- BORGES DA SILVA, E. D.; XAVIER, A.; FARIA, M. V. Joint modeling of genetics and field variation in plant breeding trials using relationship and different spatial methods: A simulation study of accuracy and bias. **Agronomy**, Basel, v. 11, n. 7, p. 1397, 2021.
- BROWN, A. V. *et al.* A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. **Nucleic Acids Research**, Oxford, v. 49, n. D1, p. D1496–D1501, 2021.
- BRUZI, A. T. **Aplicações da análise de modelos mistos em programas de seleção recorrente no feijoeiro comum**. 2008. - Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2008.
- CHOI, I.-Y. *et al.* A Soybean Transcript Map: Gene Distribution, Haplotype and Single-Nucleotide Polymorphism Analysis. **Genetics**, [s. l.], v. 176, n. 1, p. 685–696, 2007.
- COOPER, R. L. Mejoramiento por rendimiento en soya: prueba de generación temprana y adaptación específica a ambientes de alto rendimiento versus ambientes de bajo rendimiento. *In*: BEBES, S. (org.). **Temas actuales en mejoramiento del frijol**. Cali: CIAT, 1988. p. 301–323.

DE LOS CAMPOS, G.; SORENSEN, D.; GIANOLA, D. Genomic Heritability: What Is It? **PLoS Genetics**, Stanford, v. 11, n. 5, p. 1–21, 2015.

EVANGELISTA, J. S. P. C. *et al.* Environmental stratification and genotype recommendation toward the soybean ideotype: a Bayesian approach. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 21, n. 1, p. 1–8, 2021.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. Essex: Longman, 1996.

FEHR, W. R.; CAVINESS, C. E. Stages of Soybean Development. **Special report**, [s. l.], v. 80, p. 11, 1977.

FERREIRA, R. A. D. C. *et al.* Implications of the number of years assessment on recommendation of common bean cultivars. **Plant Breeding**, Berlin, v. 134, n. 5, p. 599–604, 2015.

FERREIRA, W. D. *et al.* Family number in common bean selection. **Genetics and Molecular Biology**, Ribeirão Preto, v. 23, n. 2, p. 403–409, 2000.

FOUILLOUX, G.; BANNEROT, H. Selection methods in the common bean (*Phaseolus vulgaris*). In: GEPTS, P. (org.). **Genetic resources of Phaseolus Beans: their maintenance, domestication, evolution and utilization**. Dordrecht: Kluwer Academic Publisher, 1988. p. 613.

FREY, K. J. The Use of F2 Lines in Predicting the Performance of F3 Selections in Two Barley Crosses. **Agronomy Journal**, Madison, v. 46, n. 12, p. 541–544, 1954.

GERLOFF, J. E.; SMITH, O. S. Choice of method for identifying germplasm with superior alleles - 1. Theoretical results. **Theoretical and Applied Genetics**, Berlin, v. 76, n. 2, p. 209–216, 1988.

GESTEIRA, G. de S. *et al.* Selection of early soybean inbred lines using multiple indices. **Crop Science**, Madison, v. 58, n. 6, p. 2494–2502, 2018.

GESTEIRA, G. S. *et al.* Seleção fenotípica de cultivares de soja precoce para a região Sul de Minas Gerais. **Revista Agroambiental**, Pouso Alegre, v. 7, n. 3, p. 79–88, 2015.

GLENN, K. C. *et al.* Bringing new plant varieties to market: Plant breeding and selection practices advance beneficial characteristics while minimizing unintended changes. **Crop Science**, Madison, v. 57, n. 6, p. 2906–2921, 2017.

GORNICKI, P.; FARIS, J. D. Rewiring the wheat reproductive system to harness heterosis for the next wave of yield improvement. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 111, n. 25, p. 9024–9025, 2014.

GRANT, D. *et al.* SoyBase, the USDA-ARS soybean genetics and genomics database. **Nucleic Acids Research**, Oxford, v. 38, n. SUPPL.1, p. 843–846, 2009.

HAMBLIN, J.; ZIMMERMANN, M. J. O. Breeding Common Bean for Yield in Mixtures.

**Plant Breeding Reviews**, Westport, v. 4, n. 1, p. 245–272, 1986.

HARTL, D. L.; CLARK, A. G. **Princípios de Genética de Populações**. 4. ed. Porto Alegre: Artmed, 2010.

HEDRICK, P. W. **Genetics of populations**. 4. ed. Sudbury: Jones & Bartlett Publishers, 2009.

HENDERSON, C. R. **Applications of linear models in animal breeding**. Guelph: University of Guelph, 1984.

HENDERSON, C. R. Estimation of changes in herd environment. **Journal of Dairy Science**, [s. l.], v. 32, n. 8, p. 706–706, 1949.

KERR, R. J.; DIETERS, M. J.; TIER, B. Simulation of the comparative gains from four different hybrid tree breeding strategies. **Canadian Journal of Forest Research**, Ottawa, v. 34, n. 1, p. 209–220, 2004.

KIÆR, L. P.; SKOVGAARD, I. M.; ØSTERGÅRD, H. Grain yield increase in cereal variety mixtures: A meta-analysis of field trials. **Field Crops Research**, Amsterdam, v. 114, n. 3, p. 361–373, 2009.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. ImerTest Package: Tests in Linear Mixed Effects Models. **Journal of Statistical Software**, [s. l.], v. 82, n. 13, 2017.

LEITE, M. S. O. *et al.* Sample size for full-sib family evaluation in sugarcane. **Pesquisa Agropecuária Brasileira**, Brasília, v. 44, n. 12, p. 1562–1574, 2009

LEITE, W. S. *et al.* Estimativas de parâmetros genéticos, correlações e índices de seleção para seis caracteres agrônômicos em linhagens F8 de soja. **Comunicata Scientiae**, [s. l.], v. 7, n. 3, p. 302–310, 2016.

LEMOES, R. C.; ABREU, Â. F. B.; RAMALHO, M. A. P. Procedures for identification of superior progenies in successive generations of evaluation in common bean. **Scientia Agricola**, Piracicaba, v. 77, n. 1, 2020.

LIMA, L. K.; RAMALHO, M. A. P.; ABREU, A. F. B. Implications of the progeny x environment interaction in selection index involving characteristics of the common bean. **Genetics and molecular research**, Ribeirão Preto, v. 11, n. 4, p. 4093–4099, 2012.

LUSH, J. L. Heritability of quantitative characters in farm animals. *In:* , 1949, Stockholm. (G. Bonnier & R. Larsson, Org.) **Proceedings of the 8th International Congress of Genetics**. Stockholm: Lund, 1949. p. 356–375.

MARSHALL, D. R.; BROWN, A. H. D. Stability of performance of mixtures and multilines. **Euphytica**, Wageningen, v. 22, n. 2, p. 405–412, 1973.

MARTINS, F. B. *et al.* Classificação climática de köppen e de thornthwaite para Minas Gerais: cenário atual e projeções futuras. **Revista Brasileira de Climatologia**, [s. l.], v. 22, p.

129–156, 2018.

MRODE, R. A.; THOMPSON, R. **Linear models for the prediction of animal breeding values**. 2. ed. Wallingford: CABI, 2005.

NUNES, J. A. R.; RAMALHO, M. A. P.; FERREIRA, D. F. Inclusion of genetic relationship information in the pedigree selection method using mixed models. **Genetics and Molecular Biology**, Ribeirão Preto, v. 31, n. 1, p. 73–78, 2008.

PEREIRA, F. C. *et al.* Implications of the population effect in the selection of soybean progeny. **Plant Breeding**, Berlin, v. 136, n. 5, p. 679–687, 2017.

PEREIRA, L. A.; ABREU, A. F. B.; RAMALHO, M. A. P. Variability among and within common bean progenies from segregating populations. **Annual Report of the Bean Improvement Cooperative**, Prosser, v. 60, p. 189–190, 2017.

PIEPHO, H.-P. *et al.* A stage-wise approach for the analysis of multi-environment trials. **Biometrical Journal**, Berlin, v. 54, n. 6, p. 844–860, 2012.

PIEPHO, H. P. *et al.* BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, Wageningen, v. 161, n. 1–2, p. 209–228, 2008.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2021.

RAMALHO, M. A. P. *et al.* **Aplicações da genética quantitativa no melhoramento de plantas autógamas**. Lavras: Editora UFLA, 2012.

RAMALHO, M. A. P.; DIAS, L. A. S.; CARVALHO, B. L. Contributions of plant breeding in Brazil: progress and perspectives. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 12, p. 111–120, 2012.

RAMALHO, M. A. P.; MARQUES, T. L.; LEMOS, R. C. Plant breeding in Brazil: Retrospective of the past 50 years. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 21, n. Special Issue, p. 1–11, 2021.

RAO, M. S. S. *et al.* Genotype × Environment Interactions and Yield Stability of Food-Grade Soybean Genotypes. **Agronomy Journal**, Madison, v. 94, n. 1, p. 72–80, 2002.

RESENDE, M.D.V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical (Agricultural Research in the Tropics)**, [s. l.], v. 37, n. 3, p. 182–194, 2007.

RESENDE, M. D.V. *et al.* Multigeneration index in the Within - Progenies bulk method for breeding of Self-Pollinated plants. **Crop Science**, Madison, v. 55, n. 3, p. 1202–1211, 2015.

RIBEIRO, F. O. *et al.* Genotypic Selection for Soybean Earliness. **Journal of Agricultural Science**, Ottawa, v. 12, n. 11, p. 169, 2020.

SCHMUTZ, J. *et al.* Genome sequence of the palaeopolyploid soybean. **Nature**, London, v.

463, n. 7278, p. 178–183, 2010.

SILVA, E. V. V. **Parâmetros genéticos e fenotípicos associados à precocidade em soja**. 2018. Dissertação (Mestrado em Agronomia/Fitotecnia) - Universidade Federal de Lavras, Lavras, 2018.

SILVA, G. O.; VIEIRA, J. V.; VILLELA, M. S. Tamanho de amostra para avaliação de caracteres de cenoura em sistemas de cultivo agroecológico. **Horticultura Brasileira**, Brasília, v. 27, n. 2, p. 166–170, 2009.

SILVA, H. D.; FERREIRA, D. F.; PACHECO, C. A. P. Avaliação de quatro alternativas de análise de experimentos em látice quadrado, quanto à estimação de componentes de variância. **Bragantia**, Campinas, v. 59, n. 1, p. 117–123, 2000.

SILVA, K. B. *et al.* Adaptability and stability of soybean cultivars for grain yield and seed quality. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 2, 2017.

SOARES, I. O. *et al.* Genetic and phenotypic parameters associated with soybean progenies in a recurrent selection program. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 20, n. 4, p. 1–8, 2020.

SOUZA, E. F. M.; PETERNELLI, L. A.; BARBOSA, M. H. P. Designs and model effects definitions in the initial stage of a plant breeding program. **Pesquisa Agropecuária Brasileira**, Brasília, v. 41, n. 3, p. 369–375, 2006.

TODESCHINI, M. H. *et al.* Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. **Euphytica**, Wageningen, v. 215, n. 7, 2019.

VAN EEUWIJK, F. A.; BUSTOS-KORTS, D. V.; MALOSETTI, M. What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype  $\times$  Environment Interactions? **Crop Science**, Madison, v. 56, n. 5, p. 2119–2140, 2016.

VAN GINKEL, M.; ORTIZ, R. Cross the Best with the Best, and Select the Best: HELP in Breeding Selfing Crops. **Crop Science**, Madison, v. 58, n. 1, p. 17–30, 2018.

VEIGA, R. D.; FERREIRA, D. F.; RAMALHO, M. A. P. Eficiência dos dialelos circulantes na escolha de genitores. **Pesquisa Agropecuária Brasileira**, Brasília, v. 350, n. 7, p. 1395–1406, 2000.

VOORRIPS, R. E.; MALIEPAARD, C. A. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. **BMC Bioinformatics**, [s. l.], v. 13, n. 1, p. 248, 2012.

WANG, J. *et al.* Comparison of two breeding strategies by computer simulation. **Crop Science**, Madison, v. 43, n. 5, p. 1764–1773, 2003.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. 2. ed. New York: Springer, 2016.