



**UNIVERSIDADE FEDERAL DE GOIÁS  
ESCOLA DE AGRONOMIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E  
MELHORAMENTO DE PLANTAS**

## **DIVERGÊNCIA GENÉTICA E PREDIÇÃO DE VALORES GENOTÍPICOS EM SOJA**

**CLÁUDIO ROBERTO CARDOSO DE GODOI**

Orientador:  
**Prof. João Batista Duarte**

**CLÁUDIO ROBERTO CARDOSO DE GODOI**

**DIVERGÊNCIA GENÉTICA E PREDIÇÃO DE  
VALORES GENOTÍPICOS EM SOJA**

Tese apresentada ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás, como requisito parcial à obtenção do título de Doutor em Genética e Melhoramento de Plantas.

Orientador:

**Prof. Dr. João Batista Duarte**

Goiânia, GO – Brasil

2014

**Dados Internacionais de Catalogação na Publicação (CIP)**  
**GPT/BC/UFG**

G588d Godoi, Cláudio Roberto Cardoso de.  
Divergência genética e predição de valores genotípicos em soja [manuscrito] / Cláudio Roberto Cardoso de Godoi. - 2014.  
108 f. : il, figs, tabs.

Orientador: Prof. Dr. João Batista Duarte.  
Tese (Doutorado) – Universidade Federal de Goiás,  
Escola de Agronomia, 2014.  
Bibliografia.

1. Soja – Divergência genética
2. Diversidade genética
3. Soja (*Glycine max.*) – Genótipos. I. Título.

CDU – 633.34:631.52

## TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS (TEDE) NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

**1. Identificação do material bibliográfico:**      ☐ Dissertação      ☒ Tese

### 2. Identificação da Tese ou Dissertação

|                                                                                                                     |                                                                                                                                 |        |       |
|---------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|--------|-------|
| Autor (a):                                                                                                          | Cláudio Roberto Cardoso de Godoi                                                                                                |        |       |
| E-mail:                                                                                                             | <a href="mailto:claudio_godoi@yahoo.com.br">claudio_godoi@yahoo.com.br</a>                                                      |        |       |
| Seu e-mail pode ser disponibilizado na página? <input checked="" type="checkbox"/> Sim <input type="checkbox"/> Não |                                                                                                                                 |        |       |
| Vínculo empregatício do autor                                                                                       | Nidera Sementes Ltda.                                                                                                           |        |       |
| Agência de fomento:                                                                                                 |                                                                                                                                 | Sigla: |       |
| País:                                                                                                               | Brasil                                                                                                                          | UF: GO | CNPJ: |
| Título:                                                                                                             | Divergência genética e predição de valores genotípicos em soja                                                                  |        |       |
| Palavras-chave:                                                                                                     | Glycine max, SSR, introdução de germoplasma, variabilidade genética, REML/BLUP, matriz de parentesco, modelo misto, modelo fixo |        |       |
| Título em outra língua:                                                                                             | Genetic divergence and genotypic values prediction in soybean                                                                   |        |       |
| Palavras-chave em outra língua:                                                                                     | Glycine max, SSR, germplasm introduction, genetic variability, REML/BLUP, relationship matrix, mixed model, fixed model         |        |       |
| Área de concentração:                                                                                               | Genética e melhoramento de plantas                                                                                              |        |       |
| Data defesa: (dd/mm/aaaa)                                                                                           | 07/05/2014                                                                                                                      |        |       |
| Programa de Pós-Graduação:                                                                                          | Genética e melhoramento de plantas                                                                                              |        |       |
| Orientador (a):                                                                                                     | Prof. Dr. João Batista Duarte                                                                                                   |        |       |
| E-mail:                                                                                                             | <a href="mailto:jbduarte@gmail.com">jbduarte@gmail.com</a>                                                                      |        |       |
| Co-orientador (a):*                                                                                                 |                                                                                                                                 |        |       |
| E-mail:                                                                                                             |                                                                                                                                 |        |       |

\*Necessita do CPF quando não constar no SisPG

### 3. Informações de acesso ao documento:

Concorda com a liberação total do documento    ☒ SIM      ☐ NÃO<sup>1</sup>

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF ou DOC da tese ou dissertação.

O sistema da Biblioteca Digital de Teses e Dissertações garante aos autores, que os arquivos contendo eletronicamente as teses e ou dissertações, antes de sua disponibilização, receberão procedimentos de segurança, criptografia (para não permitir cópia e extração de conteúdo, permitindo apenas impressão fraca) usando o padrão do Acrobat.

\_\_\_\_\_  
Assinatura do (a) autor (a)

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

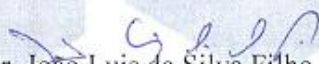
<sup>1</sup> Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

**CLÁUDIO ROBERTO CARDOSO DE GODOI**

**TÍTULO: "DIVERGÊNCIA GENÉTICA E PREDIÇÃO DE VALORES  
GENOTÍPICOS EM SOJA".**

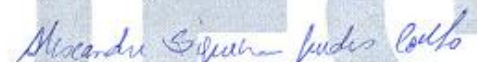
Tese DEFENDIDA em 07 de Maio de 2014, e APROVADA pela Banca  
Examinadora constituída pelos membros:

  
Prof. Dr. João Batista Duarte  
Orientador – EA/UFG

  
Dr. João Luis da Silva Filho  
Embrapa Algodão

  
Dr. José Francisco Ferraz de Toledo  
Nidera Sementes Ltda.

  
Prof. Dr. Lázaro José Chaves  
EA/UFG

  
Prof. Dr. Alexandre Siqueira Guedes Coelho  
EA/UFG

Aos meus pais Benedito Cardoso e Joana D'arc, pelo exemplo de vida, honestidade e humildade.

A minha esposa Walquiria, pelo exemplo de amor e companheirismo.

Ao meu filho Heitor, pela alegria trazida aos meus dias em sua presença.

**Dedico.**

Aos tios José Fernando e Maria Eunice.

Aos tios Antônio Cardoso (*in memoriam*) e Ilda Pereira.

Aos amigos Charles Étier e Fátima Pereira.

**Imprescindíveis** na construção dos pilares de sustentação para a minha vida acadêmica.

**Ofereço.**

## **AGRADECIMENTOS**

À Universidade Federal de Goiás, por meio da Escola de Agronomia (EA/UFG), onde tive a oportunidade de obter formação profissional, pública, gratuita e de qualidade, desde a graduação até esse instante, com o doutoramento;

Ao Programa de Pós-graduação em Genética e Melhoramento de Plantas, pela oportunidade de realização deste curso;

Aos professores Alexandre Siqueira G. Coelho, Lázaro José Chaves, Evandro Novaes, Patrícia Guimarães S. Melo, Mariana Pires C. Telles, pelos valiosos momentos de estudo, reflexão e aprendizagem no decorrer do curso;

À Nidera Sementes Ltda., em nome dos diretores Hernan Forte e José F. F. Toledo, pela confiança e a oportunidade a mim concedida com a liberação e apoio para acompanhamento do curso; também pela autorização de uso de material genético e de dados experimentais de propriedade da empresa para realização deste estudo;

Ao Professor João Batista Duarte, pela orientação, confiança e valiosos momentos de discussão e reflexão ao longo do curso;

Aos membros da Banca examinadora, pela disponibilidade em participar desta defesa e, também, pela análise crítica e valiosas sugestões de melhorias desta tese;

Ao Prof. Dr. José Baldin Pinheiro, pela importante contribuição e apoio no planejamento e execução do estudo de diversidade genética desta pesquisa;

Aos colegas Arthur Tavares O. Melo (EA/UFG), pela extração de DNA, e à Natalia S. Stabellini (ESALQ/USP), pela genotipagem, do material genético utilizado nesta pesquisa;

Aos colegas da Pesquisa Soja – Nidera Sementes Ltda., Sr. José Madalena, Wanderson Silva, Tharciso Garcia, Jesse Trindade, Cley Jeverson, Jean Carlo, Alelita Falchetti, Gilvani Matei e Luiz Antônio Cardoso Jr., pelo constante apoio e compartilhamento de atividades, de maneira a viabilizar minha participação no curso;

À minha esposa Walquiria A. Godoi e ao meu filho Heitor A. Godoi, pela credibilidade, incentivo, apoio, compreensão e paciência nos momentos de ausência, ao longo dessa jornada;

A todos que, direta ou indiretamente, acreditaram e contribuíram para a realização deste importante projeto de vida.

**“Se há sorte, eu não sei, nunca vi”**  
Renato Teixeira



## SUMÁRIO

|                                                                                                         |           |
|---------------------------------------------------------------------------------------------------------|-----------|
| <b>RESUMO GERAL .....</b>                                                                               | <b>9</b>  |
| <b>GENERAL ABSTRACT .....</b>                                                                           | <b>10</b> |
| <b>1      INTRODUÇÃO GERAL .....</b>                                                                    | <b>11</b> |
| <b>2      REVISÃO BIBLIOGRÁFICA .....</b>                                                               | <b>15</b> |
| 2.1    ORIGEM, BASE GENÉTICA E DIVERSIDADE DA SOJA CULTIVADA .....                                      | 15        |
| 2.1.1    Base genética e melhoramento na espécie <i>Glycine max</i> .....                               | 16        |
| 2.1.2    Diversidade genética e sua mensuração no melhoramento da soja .....                            | 19        |
| 2.2    MODELOS LINEARES MISTOS .....                                                                    | 21        |
| 2.2.1    Modelagem estatística em modelos lineares mistos .....                                         | 23        |
| 2.2.2    Matrizes <i>G</i> e <i>R</i> em modelos lineares mistos .....                                  | 24        |
| 2.2.3    Propriedades e vantagens da metodologia de predição REML/BLUP .....                            | 28        |
| 2.2.4    Critérios de escolha e seleção de modelos de análise .....                                     | 32        |
| <b>3      DIVERSIDADE E ESTRUTURA GENÉTICA EM GERMOPLASMA DE SOJA .....</b>                             | <b>34</b> |
| RESUMO .....                                                                                            | 34        |
| ABSTRACT .....                                                                                          | 34        |
| 3.1    INTRODUÇÃO .....                                                                                 | 35        |
| 3.2    MATERIAL E MÉTODOS .....                                                                         | 38        |
| 3.2.1    Material experimental .....                                                                    | 38        |
| 3.2.2    Extração de DNA e genotipagem .....                                                            | 39        |
| 3.2.3    Análise genética e populacional .....                                                          | 40        |
| 3.3    RESULTADOS E DISCUSSÃO .....                                                                     | 42        |
| 3.3.1    Análise genética descritiva e da diversidade genética .....                                    | 42        |
| 3.3.2    Divergência genética e estruturação populacional do germoplasma .....                          | 46        |
| 3.4    CONCLUSÕES .....                                                                                 | 57        |
| <b>4      PREDIÇÃO DE VALORES GENOTÍPICOS EM ENSAIOS DE COMPETIÇÃO FINAL DE GENÓTIPOS DE SOJA .....</b> | <b>59</b> |
| RESUMO .....                                                                                            | 59        |
| ABSTRACT .....                                                                                          | 60        |
| 4.1    INTRODUÇÃO .....                                                                                 | 60        |
| 4.2    MATERIAL E MÉTODOS .....                                                                         | 65        |
| 4.2.1    Caracterização experimental .....                                                              | 65        |
| 4.2.2    Análises biométricas .....                                                                     | 67        |
| 4.2.2.1    Estimação de efeitos genotípicos por modelo fixo .....                                       | 68        |
| 4.2.2.2    Predição de valores genotípicos por modelos mistos .....                                     | 69        |
| 4.2.2.3    Flexibilização da matriz <i>R</i> para predição de valores genotípicos .....                 | 72        |
| 4.2.3    Comparações entre abordagens estatísticas .....                                                | 73        |
| 4.2.4    Simulação de redução de repetições e desbalanceamento nos dados .....                          | 75        |
| 4.2.5    Aplicativos computacionais empregados nas análises .....                                       | 75        |
| 4.3    RESULTADOS E DISCUSSÃO .....                                                                     | 76        |
| 4.3.1    Avaliação do ajuste e seleção dos modelos de análise .....                                     | 76        |

|       |                                                                                         |     |
|-------|-----------------------------------------------------------------------------------------|-----|
| 4.3.2 | <b>Influência da modelagem em parâmetros genéticos e na seleção genotípica .....</b>    | 81  |
| 4.3.3 | <b>Efeito da redução de repetições e do desbalanceamento na escolha do modelo .....</b> | 93  |
| 4.4   | <b>CONCLUSÕES .....</b>                                                                 | 95  |
| 5     | <b>CONCLUSÕES GERAIS .....</b>                                                          | 96  |
| 6     | <b>REFERÊNCIAS .....</b>                                                                | 98  |
|       | <b>APÊNDICES .....</b>                                                                  | 107 |

## RESUMO GERAL

GODOI, C. R. C. **Divergência genética e predição de valores genotípicos em soja**. 2014. 108 f. Tese (Doutorado em Genética e Melhoramento de Plantas)–Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.<sup>2</sup>

Os programas de melhoramento de soja visam à seleção de genótipos de alto valor genético, com a finalidade de uso principalmente em duas de suas etapas: a) como genitores no processo de hibridação (fase inicial); e, b) para indicação como nova cultivar (fase final). Nesse contexto, num primeiro estudo avaliou-se, por meio de marcadores microsatélites (SSR), a diversidade genética em germoplasma de soja adaptado às condições brasileiras. O material experimental constituiu-se de 192 acessos, entre introduções e germoplasma de origem nacional. Na avaliação da divergência genética, considerou-se a análise descritiva e a distância genética de Rogers-W. Nos 37 locos genotipados, identificaram-se 222 alelos, com média de seis alelos por loco e variação de 2 a 14 alelos. O agrupamento dos genótipos mostrou-se associado à origem do germoplasma, resultando em dois grupos: um introduzido e outro brasileiro. Das estimativas de distâncias genéticas obtidas, 85% foram superiores a 0,70, indicando bom potencial do germoplasma para hibridações em programas de melhoramento da soja. Concluiu-se que os marcadores SSR são úteis na identificação de grupos genotípicos divergentes, bem como de combinações de alta variabilidade genética. Ademais, o uso de germoplasma introduzido garante a incorporação de alelos necessários à ampliação da base genética da espécie e, consequentemente, da variabilidade necessária para uso no processo seletivo. Num segundo estudo, no contexto da análise de modelos mistos, avaliaram-se estratégias de estimação e predição de valores genotípicos para produtividade de grãos, a partir de ensaios de competição final de linhagens de soja. Os genótipos, em número de 111 e classificados em três grupos de maturação, foram semeados em até 23 experimentos conduzidos na região central do Brasil. Os experimentos foram conduzidos no delineamento de blocos completos casualizados, com três repetições. Nas análises biométricas adotaram-se as abordagens de modelo fixo e de modelo misto, neste caso, assumindo-se efeitos genotípicos como aleatórios. Na última abordagem, consideraram-se ainda análises com ou sem uso da informação de parentesco genético, obtida a partir de genealogias ou por marcadores SSR, e associada à matriz de covariâncias dos efeitos aleatórios ( $G$ ). Para cada modelo, num contexto de análise espacial, adotaram-se também distintas estruturas para a matriz de covariâncias residuais ( $R$ ). Concluiu-se, então, que: i) a análise com modelo fixo é adequada para estimar efeitos genotípicos em soja, sob condições de balanceamento dos dados e ortogonalidade do delineamento; ii) sob tais condições, a inclusão da informação de parentesco associada à matriz  $G$ , embora não garanta melhor ajuste aos modelos, sob herdabilidade moderada ou baixa, melhora a precisão das predições de valores genotípicos; iii) o uso de estruturas espaciais associadas à matriz  $R$ , na presença de autocorrelação residual, melhora o ajuste estatístico dos modelos; e, iv) corrobora-se a tese de que, sob alta herdabilidade, a escolha do modelo de análise não altera o posicionamento relativo dos genótipos, e, portanto, não impacta significativamente na seleção de genótipos superiores.

**Palavras-chave:** *Glycine max*, SSR, introdução de germoplasma, variabilidade genética, REML/BLUP, matriz de parentesco, modelo misto, modelo fixo.

---

<sup>2</sup> Orientador: Prof. Dr. João Batista Duarte. EA-UFG.

## GENERAL ABSTRACT

GODOI, C. R. C. **Genetic divergence and genotypic values prediction in soybean.** 2014. 108 f. Thesis (Doctorate in Genetic and Plant Breeding)—Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.<sup>3</sup>

Soybean breeding programs practice selection of high genetic value genotypes with two main objectives: a) to use them as parents in the hybridization process (first stage of the program), and b) to indicate them as new cultivars (final stage of the program). In this context, a first study used microsatellite markers (SSR) to assess the genetic diversity of soybean germplasm adapted to the Brazilian conditions. The experimental material consisted of 192 accessions, which included both introductions and Brazilian germplasm. The genetic divergence was assessed by descriptive analysis and the Rogers-W genetic distance. A total of 222 alleles were identified in the 37 genotyped loci, with an average of six alleles and a range of 2 to 14 alleles per locus. The genotypes were clustered according to the origin of the germplasm, and resulted in two groups: one group formed by introductions and other by Brazilian genotypes. Eighty five percent of the genetic distances estimates were above 0.70, suggesting that the assessed germplasm has good potential for hybridization in soybean breeding programs. It was concluded that the SSR markers are useful to identify divergent genotypic groups, as well as genotypic combinations with high genetic variability. It also became clear that the use of introduced germplasm ensures the incorporation of alleles necessary to increase the genetic base of soybeans and, consequently, the variability needed for the selective process. In a second study, the mixed model approach was used to assess some strategies of estimation and prediction of genotypic values for grain yield in the soybean regional yield trials. A total of 111 genotypes classified into three maturity groups were sown in up to 23 experiments in Central Brazil. The experiments were carried out in randomized complete block designs, with three replications. The biometrical analyses followed the fixed model and mixed model approaches, in the latter case assuming the genotypic effects as random. In the mixed model approach, analyses were made with or without information from the relationship estimates obtained either by genealogy or SSR markers, arranged in a genotypic covariance matrix ( $G$ ). Also, in a context of spatial analysis, different structures were used in the residual covariance matrix ( $R$ ) for each mixed model adjusted. The following conclusions were obtained: i) the fixed model analysis is adequate to estimate genotypic values in soybean trials with balanced data and orthogonal design; ii) under such conditions and intermediate to low heritability, the inclusion of relationship information associated to  $G$  matrix, although does not ensure the best fit models, improves the precision in predicting genotypic values; iii) the use of spatial structures associated to  $R$  matrix, in presence of the residual autocorrelation, improves the goodness of model fit to the data; and, iv) the choice of model for the analysis does not change the ranking of the genotypes in high heritability situations and, therefore, does not impact significantly on the selection of superior genotypes.

**Key words:** *Glycine max*, SSR, germplasm introduction, genetic variability, REML/BLUP, relationship matrix, mixed model, fixed model.

---

<sup>3</sup> Adviser: Prof. Dr. João Batista Duarte. EA-UFG.

## 1 INTRODUÇÃO GERAL

A soja (*Glycine max* (L.) Merrill) está entre as culturas oleaginosas de maior importância mundial. Dentre os seus usos, destaca-se a extração de subprodutos como o farelo e o óleo, sendo que de sua proteína são obtidos diversos produtos comestíveis; além da soja integral que é utilizada na indústria de alimentos (Sediyama et al., 1999). Atualmente, o Brasil é o segundo maior produtor mundial e o principal exportador dessa oleaginosa. Na safra atual (2013/14), estima-se que o país atinja uma produção de 86 milhões de toneladas de soja, cultivada em uma área de 30 milhões de hectares (Conab, 2014).

Apesar de frequentes relatos de que a espécie possui base genética estreita (Hiromoto & Vello, 1986; Gizlice et al., 1994; Priolli et al., 2002; Mikel et al., 2010; Wysmierski, 2010), os programas de melhoramento têm obtido sucesso na seleção de cultivares. Os ganhos genéticos em produtividade de grãos, nessa cultura, encontram-se entre 1,2% e 2,0% ao ano (Toledo et al., 1990; Ferrarotti, 2004; Rossi, 2004; Toledo et al., 2004). Apesar disso, para a continuidade desses ganhos tornam-se necessários o conhecimento e a adoção de estratégias racionais de utilização do germoplasma disponível. Só assim é possível garantir alto valor genético e divergência suficiente para uso posterior nos processos de hibridação, geração de populações segregantes e seleção de genótipos superiores para os caracteres de interesse.

Vários autores relatam que a introdução de germoplasma, sobretudo proveniente de *pools* gênicos distintos, é uma alternativa para aumentar a diversidade genética disponível (Hiromoto & Vello, 1986; Priolli et al., 2004; Yamanaka et al., 2007; Mikel et al., 2010; Mulato et al., 2010; Wysmierski, 2010). Para a quantificação dessa diversidade, bem como do potencial de uso do germoplasma, os marcadores moleculares, especialmente os microssatélites (SSR), tem-se constituído em ferramenta de grande aplicação e eficiência (Ferreira & Grattapaglia, 1998; Garcia et al., 2007; Hwang et al., 2009; Priolli et al., 2010). Entre as aplicações da informação gerada por esse tipo de marcador, destaca-se o suporte objetivo dado à escolha de genitores, para fins de hibridação dirigida (Falconer & Mackay, 1996; Ferreira & Grattapaglia, 1998; Ramalho et

al., 2001). Também merece destaque a possibilidade de uso dessa informação para melhorar a acurácia de estimativas ou predições de valor genético de indivíduos sujeitos ao processo de seleção (Bernardo, 1994; André, 1999; Silva Filho, 2004; Bauer et al., 2006; Nunes et al., 2008; Piepho et al., 2008; Kelly et al., 2009; Lorenzana & Bernardo, 2009; Pinheiro et al., 2013).

Nesse contexto, os programas de melhoramento de plantas visam à identificação de genótipos com desempenho superior, seja para uso como potenciais genitores em programas de hibridação, seja para recomendação como novas cultivares. Para isso, torna-se necessária a avaliação prévia dos genótipos candidatos à seleção, em experimentos de campo, os quais devem ser conduzidos em vários ambientes dentro da região alvo de recomendação. O objetivo central desse tipo de avaliação é inferir sobre os seus valores genéticos e, conseqüentemente, sobre seu respectivo ordenamento para fins de seleção (Resende, 2007).

Quando se pensa na avaliação experimental, esta deve ser a mais eficiente e precisa possível, de forma a otimizar os investimentos e recursos financeiros direcionados a essa atividade. Além disso, com o avanço do processo de melhoramento genético das espécies cultivadas, as diferenças entre as cultivares ficam cada vez menores, tornando-se necessário o uso de técnicas e procedimentos experimentais eficientes para detectar essas pequenas diferenças (Ramalho et al., 2000). Nesse sentido, torna-se de primordial importância a escolha adequada do método de estimação, com o intuito de propiciar inferências mais precisas e realistas sobre o potencial dos tratamentos genéticos. Conforme Smith et al. (2005), é imprescindível que os métodos biométricos usados para experimentação e análise de dados, em programas de melhoramento de plantas, sejam de boa acurácia, eficiência e informativos. Resende (2007) recomenda que, na escolha de modelos de estimação, deve-se optar por aqueles que apresentem boa acurácia seletiva, ausência de vício, precisão e menor erro quadrático médio.

Considerando-se o estudo de caracteres quantitativos, geralmente, o fenótipo do indivíduo não reflete fielmente o seu valor genotípico. Nesse caso, torna-se de grande importância o uso de métodos que otimizem o uso da informação disponível, no sentido de gerar uma classificação de progênes mais próxima possível do verdadeiro ordenamento dos valores genotípicos (White & Hodge, 1989). Nesse sentido, torna-se importante também a avaliação das abordagens tradicionais, nas quais grande número de melhoristas se apoia para o processo de seleção. É comum, por exemplo, a seleção fenotípica baseada

em médias, obtidas por métodos de estimação que assumem os efeitos de tratamentos como fixos. Nessa abordagem, além de se considerar a suposição pouco realista de que o coeficiente de herdabilidade é máximo ( $h^2 = 1,0$ ) para o caráter sob seleção, há ainda a prerrogativa de que esse procedimento apenas propicie inferências sobre os fenótipos observados, numa abordagem *posdictiva* (Resende, 2007).

De outra forma, é interessante considerar a seleção com base nas médias genotípicas (ou valores genotípicos) dos tratamentos, ao contrário das médias fenotípicas. Ou seja, focar a seleção genética para o caráter de interesse, considerando os valores genotípicos dos indivíduos sob seleção. Nessa abordagem, consideram-se, na estimação, os tratamentos genéticos como sendo de efeito aleatório. Com isso, adota-se uma suposição mais realista, principalmente, sob condições em que ocorre relacionamento genético entre esses tratamentos, além da necessidade de lidar com caracteres quantitativos. A adoção de efeitos de tratamentos como aleatórios flexibiliza mais realisticamente a condição dos valores de herdabilidade ( $0 \leq h^2 \leq 1,0$ ), proporcionando o ajuste nos valores fenotípicos por uma função da herdabilidade do caráter sob seleção. Isto tem, como consequência, a possível eliminação dos efeitos residuais de ambiente embutidos nos dados fenotípicos (efeito de “encolhimento” ou *shrinkage*), além de outros benefícios (Resende, 2007).

Nos últimos anos, grande volume de pesquisas tem sido gerado, ressaltando a importância do uso dessa abordagem na predição de valores genéticos; principalmente, devido à sua flexibilidade de aplicação e as propriedades estatísticas desejáveis de seus estimadores e, ou, preditores (André, 1999; Duarte & Vencovsky, 2001; Resende, 2007; Piepho et al., 2008; Nunes et al., 2008; Felipe & Duarte, 2010). Segundo Piepho et al. (2008), em culturas anuais a metodologia de modelos mistos (BLUP) não é tão importante porque, usualmente, se produz uma grande quantidade de dados fenotípicos por genótipo, os quais são, ainda, conduzidos numa série de experimentos. Assim, os diferentes métodos de estimação nem sempre fornecem resultados muito discrepantes, fornecendo a falsa impressão de que a escolha do estimador não é importante.

Na presente pesquisa objetivou-se avaliar, por meio de marcadores microsatélites (SSR), a diversidade genética em germoplasma de soja adaptado às condições brasileiras, com vista ao dimensionamento de seu potencial de uso. Adicionalmente, no contexto da análise de modelos mistos, avaliaram-se estratégias de estimação ou predição de valores genotípicos, para produtividade de grãos, a partir de ensaios de competição final de linhagens de soja. Para isso, fez-se uso de informações

geralmente disponíveis num programa de melhoramento genético, como os coeficientes de parentesco dos genótipos, obtidos diretamente de genealogias ou estimados via marcadores moleculares. Em síntese, buscou-se identificar alternativas que otimizem o uso da informação associada aos dados, visando a adoção de estratégias mais eficientes e eficazes para subsidiar o processo de seleção nos programas de melhoramento de soja.



## 2 REVISÃO BIBLIOGRÁFICA

### 2.1 ORIGEM, BASE GENÉTICA E DIVERSIDADE DA SOJA CULTIVADA

A soja é uma planta de clima temperado, com ampla adaptação nos climas subtropicais e tropicais (Sediyama et al., 1999). Botanicamente é classificada na classe Dicotyledoneae, família Fabaceae, gênero *Glycine* (L.), sub-gênero *Glycine* subg. soja (Moench), espécie *Glycine max* (L.) Merrill (Gazzoni, 1995). Esta espécie possui número diploide de cromossomos  $2n=40$  (Hymowitz, 2004), com genoma de 950 Mbp sequenciado (Schmutz et al., 2010). As flores da soja são hermafroditas completas e suas plantas se reproduzem tipicamente por meio de autopolinização. Portanto, é uma planta tipicamente autógama, apresentando, em média, uma taxa de cruzamentos naturais inferior a 1% (Gazzoni, 1995; Sediyama et al., 1999).

A soja cultivada (*G. max*) é frequentemente relacionada como sendo originária do Nordeste da China, entre as latitudes 35° e 45° N (Hymowitz, 2004). Contudo, estudos recentes indicam que o evento de domesticação ocorreu por um evento único no Sul da China, advindo da espécie *Glycine soja*, que ocorre na forma selvagem na região leste asiática (Guo et al., 2010).

No processo de domesticação, as mudanças evolutivas principais foram: aumento do tamanho das sementes, mudança na forma e cor das sementes, alteração na forma e tamanho das folhas, hábito de crescimento ereto, ausência de deiscência das vagens, ocorrência de flores brancas e pubescência cinza (Wang & Li, 2011). Também houve perda de, aproximadamente, 50% da diversidade genética, principalmente de alelos raros (Hyten et al., 2006; Guo et al., 2010). Contudo, diversos estudos indicam que no *pool* gênico primário encontra-se variabilidade genética disponível suficiente para atividades de melhoramento da espécie, considerando o uso de variedades crioulas (*landraces* ou *plant introductions* – *PI*), cultivares melhoradas (obsoletas e modernas) ou mesmo acessos da espécie selvagem *G. soja* (Hymowitz, 2005; Hyten et al., 2006; Yamanaka et al., 2007).

### 2.1.1 Base genética e melhoramento na espécie *Glycine max*

A base genética de uma espécie cultivada pode ser definida como sendo o número de linhagens ancestrais envolvido na formação de um grupo representativo de cultivares, associado à contribuição genética relativa destas linhagens na formação de cada cultivar (Wysmierschi, 2010). Esse conceito representa uma medida de diversidade genética do germoplasma disponível ao melhoramento da espécie, que permite comparar, de maneira relativa, diferentes programas de melhoramento genético.

Apesar da grande importância econômica da soja no mundo, esta espécie cultivada é comumente reportada na literatura como sendo de base genética estreita. Wysmierschi (2010), ao comparar a base genética e o coeficiente médio de parentesco da soja, entre alguns países (Estados Unidos, China, Japão, Índia e Brasil), relata que a China representa a maior base genética, enquanto o Brasil possui a base genética mais estreita. Dentre as razões para esse estreitamento da base genética destaca-se o baixo nível de variabilidade genética na própria espécie selvagem, *G. soja*, se comparada a de outras espécies vegetais (Hyten et al., 2007). Acrescentam o afunilamento da diversidade genética (*bottleneck*) ocorrido durante a sua domesticação, e, recentemente, devido ao processo de melhoramento genético (Hyten et al., 2007; Guo et al., 2010). O melhoramento tem contribuído para esse efeito, sobretudo pelo pequeno número de genótipos ancestrais (*landraces* ou *plant introductions* – *PI*) que foram introduzidos dos centros de origem da espécie para a formação da base inicial dos principais programas de melhoramento no mundo. Atualmente, esses programas usam como genitores apenas um grupo limitado de genótipos elite, basicamente cultivares de alta “performance” (Hiromoto & Vello, 1986; Gizlice et al., 1994; Hyten et al., 2007; Mikel et al., 2010; Wysmierschi, 2010).

Com a domesticação, ocorreu a maior perda de diversidade genética na soja (Hyten et al., 2006; Guo et al., 2010). Esses estudos indicam redução de 50% da diversidade da espécie e perda de 81% de alelos raros (com frequência inferior a 0,05). Guo et al. (2010) verificaram que a diversidade genética em *landraces* foi equivalente à metade da encontrada na espécie selvagem, sugerindo forte afunilamento da diversidade. Estes autores observaram drástica redução na heterozigosidade esperada, considerando a condição de equilíbrio de Hardy-Weinberg para os locos estudados, tendo detectado apenas 41,9% da riqueza alélica (número de alelos por loco) da espécie selvagem nas *landraces* estudadas. Hyten et al. (2006) avaliaram a diversidade genética em genótipos selvagens,

*landraces* e cultivares melhoradas americanas (obsoletas e modernas), tendo constatado que as cultivares modernas retiveram 72% da diversidade presente em *landraces* de origem asiática; contudo, houve perda de 79% de alelos raros (frequência inferior a 0,10). Eles relacionaram a perda de diversidade ao pequeno número de introduções asiáticas na formação inicial dos programas de melhoramento, nos Estados Unidos, e não à seleção artificial imposta no processo de melhoramento. Isso porque não detectaram diferença expressiva entre as diversidades de cultivares ancestrais (antigas e obsoletas) e cultivares elite (modernas).

Gizlice et al. (1994) observaram que cerca de 75% dos genes das cultivares modernas, nos Estados Unidos, provêm das dezessete primeiras linhagens desenvolvidas no país, antes de 1960. Caracterizaram, ainda, que 95% dos genes encontrados em cultivares modernas foram provenientes de 35 genótipos, dos quais 28 são ancestrais (*PI* ou *landraces*) e sete, provenientes de melhoramento. Nesse grupo, quatro cultivares (Lincoln, Harosoy, Lee e D49-2491) contribuíram com cerca de 40% dos genes das cultivares norte americanas modernas. Também nos EUA, Mikel et al. (2010) verificaram que, entre os anos de 1970 e 2008, dez cultivares representaram 16,8% dos genitores usados no desenvolvimento de novas cultivares. No período de 1999 a 2008, constataram que 23% tiveram contribuição mínima de 25% de um único genitor (cultivar A3127). Isso caracteriza o referido estreitamento da base genética em uso no melhoramento da espécie; isto é, poucos genótipos como fonte de germoplasma na geração da maioria das cultivares modernas.

Hiromoto & Vello (1986) constataram que as cultivares brasileiras, até a década de 1980, foram derivadas basicamente de 26 introduções norte americanas. Destas, apenas quatro representavam cerca de 50% do *pool* gênico brasileiro; ou, de outro modo, nove foram responsáveis por mais de 80% do conjunto de genótipos amostrado. Priolli et al. (2002), avaliando um grupo de 186 cultivares brasileiras, identificaram que, do agrupamento obtido com uso de doze marcadores microsatélites, as cultivares norte americanas usadas como genitores (Santa Rosa, D49-2491, Hill, Davis e Hood) tiveram maior contribuição efetiva na formação desse grupo de germoplasma. Em estudo mais recente, visando conhecer a base genética de 444 cultivares brasileiras, de origem pública e privada, Wysmierski (2010) verificou a presença de sessenta ancestrais comuns na genealogia destas cultivares. Destes, quatorze linhagens ancestrais representam 92,41% da base genética brasileira, e, destas, apenas quatro ancestrais (CNS, S-100, Nanking e

Tokyo) contribuíram com 55,26%. Assim, o autor conclui que a base genética atual da soja brasileira não sofreu mudanças expressivas desde 1986, permanecendo similar à base genética da soja nos EUA, com a qual compartilha seis das principais linhagens ancestrais. Este resultado é coerente ao encontrado por Miranda et al. (2007), que constataram a manutenção do tamanho efetivo populacional ( $N_e$ ) da soja brasileira ao longo das últimas décadas.

A grande contribuição dos genótipos CNS e S-100 à base genética da soja brasileira decorre, basicamente, do fato de que, do cruzamento destes ancestrais foram selecionadas a cultivar Lee e a linhagem D49-2491, que é genitor da cultivar Bragg. As cultivares Bragg e Lee representam, respectivamente, 23,65% e 27,05% dessa base genética, seja por seu grande uso direto ou pelo de seus descendentes, nos cruzamentos inicialmente realizados no Brasil (Wysmierski, 2010).

Priolli et al. (2004) constataram que a variabilidade genética da soja no Brasil se manteve ao longo de três décadas. Notaram a preservação de alelos de cultivares antigas (Santa Rosa, IAC-8, FT-Cristalina, Paraná) nas cultivares modernas, contudo, alegaram que um dos fatores para a manutenção da variabilidade foi a introdução de novos materiais genéticos (Centennial, Stonewall, Forrest e algumas *PI*). Os autores estudaram o efeito do melhoramento sobre a diversidade genética em seis programas nacionais, tendo observado que alguns locos eram compostos de alelos exclusivos, associados a alguns programas, enquanto outros por alelos comuns, de maior frequência em todos os programas. Na comparação entre períodos, não observaram diferenças significativas na variabilidade total. Logo, concluíram que o germoplasma brasileiro de soja manteve nível constante de diversidade genética no período estudado, além de heterogeneidade entre determinados programas.

Mesmo com os diversos relatos de reduzida base genética, os programas de melhoramento de soja têm obtido sucesso na seleção de cultivares com alto potencial produtivo. Nos Estados Unidos, o ganho genético médio foi de 3,2% por ciclo de melhoramento (Mikel et al., 2010). Na Argentina esse ganho, em produtividade de grãos, foi de 1,2% a 2,0% ao ano (Ferrarotti, 2004; Rossi, 2004). No Brasil, considerando apenas o Estado do Paraná, os ganhos genéticos entre os anos de 1981 e 1986 foram de 1,8% e 1,3%, nos grupos de maturação precoce e tardio, respectivamente (Toledo et al., 1990). Em outro estudo, Toledo et al. (2004) concluíram que os programas brasileiros de melhoramento de soja foram eficientes no desenvolvimento de cultivares, tendo

incrementado ganhos genéticos em produtividade de grãos na magnitude de 1,5% a 2,0% ao ano.

Dentre as estratégias que visam a manutenção dos ganhos genéticos, Yamanaka et al. (2007) indicam que uma alternativa para aumentar essa diversidade seria a introdução de cultivares ou linhagens elite com boas características agronômicas, provenientes de diferentes *pools* gênicos, pois isto facilita a incorporação direta em programas de melhoramento. Adicionalmente, a introdução dirigida de linhagens (*landraces* ou *PI*) permite a incorporação de alelos raros e, ou, exclusivos, não encontrados ou disponíveis nas coleções de germoplasma e cultivares elite. Isto gera benefícios diretos aos programas de seleção, principalmente pela incorporação de alelos para características específicas como resistência a doenças e nematóides (Hyten et al., 2006; Mulato et al., 2010). Miranda et al. (2007) relatam que nos últimos anos a introdução de germoplasma de soja no Brasil enfocou, principalmente, as fontes de resistência ao cancro da haste e ao nematóide de cisto, tendo contribuído para a manutenção da variabilidade genética nas cultivares brasileiras.

### **2.1.2 Diversidade genética e sua mensuração no melhoramento da soja**

O conhecimento da variabilidade genética disponível e sua estrutura são de grande importância para o planejamento de hibridações num programa de melhoramento de plantas; pois, possibilita a obtenção de populações segregantes com bons níveis de variabilidade genética e alto valor adaptativo para os caracteres de interesse. Nessa fase, além de preconizar a combinação de materiais genéticos superiores, devem-se utilizar genótipos que possuam maior distância genética entre si, visando, com isso, a maximização das chances de obtenção de novas combinações gênicas favoráveis (Falconer & Mackay, 1996; Ferreira & Grattapaglia, 1998; Ramalho et al., 2001). Além disso, o uso de genitores divergentes geneticamente assegura medidas de proteção contra problemas futuros advindos da uniformidade genética na cultura, como maior vulnerabilidade a pragas ou doenças, que podem ser provocados pelo uso extensivo de uma ou mais cultivares estreitamente relacionadas (Vieira et al., 2009).

O melhoramento genético, em espécies cultivadas como a soja, tem proporcionado redução na diversidade genética, devido ao uso de pequeno grupo de germoplasma elite, formado basicamente por cultivares comerciais, para o

desenvolvimento de novas cultivares (Mikel et al., 2010). Nesse sentido, a introdução de germoplasma (linhagens exóticas ou cultivares e linhagens elite melhoradas), principalmente oriundo de *pools* gênicos distintos, permite o incremento da variabilidade disponível para uso no processo de melhoramento, evitando o estreitamento da base genética e, conseqüentemente, uma possível limitação dos ganhos genéticos nos sucessivos ciclos de melhoramento (Hiromoto & Vello, 1986; Priolli et al., 2004; Yamanaka et al., 2007; Mikel et al., 2010; Mulato et al., 2010; Wysmierski, 2010). Logo, essa estratégia fornece a base para a manutenção dos ganhos genéticos nos programas de seleção.

Dentre outros métodos, o uso de marcadores moleculares tem-se constituído em ferramenta de grande aplicação e eficiência na avaliação da diversidade genética. Entre estes, destacam-se os microssatélites, amplamente utilizados por sua elevada capacidade de detecção de polimorfismo, principalmente por seu caráter multialélico (Ferreira & Grattapaglia, 1998). Essa classe de marcadores baseia-se na variação do comprimento de regiões de DNA repetitivo, sendo amplificados por PCR, com *primers* específicos de sequências conhecidas de nucleotídeos (de vinte a trinta pares de bases), flanqueando a região de interesse (Souza, 2007).

Conforme descrito por Ferreira & Grattapaglia (1998) e Souza (2007), os microssatélites (ou SSR, sequências simples repetidas) consistem de pequenas sequências contendo entre um e seis nucleotídeos de comprimento, repetidas lado a lado, denominadas *motives*. Essas sequências são amplamente distribuídas no genoma, geralmente ocorrendo a cada intervalo de 50 Kbp e 100 Kbp, sendo que, em plantas, os *motives* mais comuns são formados por dois nucleotídeos do tipo AT, GA e GT. Em eucariotos, de forma geral, essas sequências simples são frequentes e distribuídas ao acaso no genoma, formando locos genéticos altamente polimórficos, multialélicos e de grande conteúdo informativo do ponto de vista genético. Em geral, estas regiões são hipervariáveis e a variabilidade detectada é bastante elevada, o que torna esse marcador com característica multialélica. Segundo Morgante et al. (2002), o polimorfismo detectado por estes marcadores situa-se, quase que exclusivamente, em regiões intergênicas (ou não codificantes), o que torna estes marcadores teoricamente neutros do ponto de vista da ação da seleção, seja natural ou artificial.

Hwang et al. (2009) e Priolli et al. (2010) enumeram algumas características desses marcadores, as quais têm favorecido o seu emprego na estimação da diversidade genética em soja: abundância no genoma; alto nível de polimorfismo; codominância;

detecção por PCR; posição conhecida no genoma; e capacidade de automatização da leitura. Tudo isso favorece a identificação fácil e precisa de diferentes alelos por loco. Ademais, são altamente efetivos na distinção de acessos com estreita base genética, como é o caso do germoplasma brasileiro de soja (Priolli et al., 2010). Garcia et al. (2007) indicaram o uso destes marcadores, pois possibilitam maior detecção de polimorfismo de DNA, contribuindo para a identificação e discriminação de cultivares de soja, com maior precisão e acurácia.

## 2.2 MODELOS LINEARES MISTOS

A variação entre dados pode ser estudada considerando-se diferentes classes num mesmo modelo linear. Assim, modelos em que os efeitos são apenas fixos são chamados de modelos de efeitos fixos; modelos que contêm efeitos fixos e aleatórios são denominados de modelos mistos; e aqueles que contêm apenas efeitos aleatórios são chamados de modelos aleatórios. Portanto, quando parâmetros que compõem o modelo compreendem uma mistura de fatores de natureza fixa e aleatória, este é definido como um modelo misto (Searle et al., 1992). Segundo estes autores, embora não seja considerado na prática, a rigor, todo modelo que inclua a média é um modelo misto, porque também sempre contém o erro ou resíduo, aleatório; caracterizando-se, assim, uma mistura de fatores de efeitos fixo e aleatório. Duarte (2000) acrescenta que a modelagem mista generaliza qualquer estrutura de correlação entre as observações; logo, os outros dois tipos (fixos e aleatórios) representam apenas casos particulares de modelos mistos.

No contexto dos modelos mistos, o procedimento BLUP (“Best Linear Unbiased Prediction” – melhor predição linear não viciada) foi desenvolvido por Henderson, em 1949, com a finalidade de predizer valores genéticos de animais, para produção de leite, sob condições de desbalanceamento nos dados (White & Hodge, 1989). Nesse método são obtidos, simultaneamente, num mesmo modelo, os estimadores BLUE (“Best Linear Unbiased Estimator” – melhor estimador linear não viciado) para efeitos fixos, e os preditores BLUP, para os valores genéticos, efeitos aleatórios (Henderson, 1984; Mrode, 1996).

Define-se BLUP como um método padrão para predizer efeitos aleatórios de um modelo misto (Piepho et al., 2008), sendo considerado um procedimento ótimo de seleção para efeitos genéticos e genotípicos (Resende, 2007). Também é um procedimento

geral que permite comparações entre genótipos de diferentes populações (genitores) e avaliados em diferentes conjuntos de ambientes (Bernardo, 2002).

Desde a sua proposição, o procedimento BLUP tem sido amplamente usado em avaliação genética de animais domésticos, por suas propriedades estatísticas. Inicialmente, foi usado na aplicação de modelos simples como o modelo animal, especialmente em bovinos, em que se lida com a situação em que não se consegue observar diretamente o efeito fenotípico, por exemplo, na obtenção do valor genético do reprodutor na produção de leite em gado leiteiro (Mrode, 1996). Contudo, com o desenvolvimento e disponibilização de recursos computacionais, possibilitou-se seu uso associado a modelos mais complexos, além da aplicação no melhoramento de plantas.

No melhoramento de plantas, em geral, a abordagem de modelos mistos tem menor aceitação. Contudo, é comum seu uso no melhoramento de plantas perenes e florestais (White & Hodge, 1989; Bernardo, 2002; Resende, 2002, 2007). Em culturas anuais há menor adoção da metodologia, sendo que só após a década de 1990 houve iniciativas mais concretas de uso de BLUP (Bernardo, 1994; André, 1999; Bernardo, 2002; Resende, 2007).

Nos últimos anos, grande volume de pesquisas tem sido gerado, ressaltando a importância de uso dessa abordagem na predição de valores genéticos ou genotípicos, principalmente, devido à sua flexibilidade de aplicação e às propriedades estatísticas desejáveis (André, 1999; Duarte & Vencovsky, 2001; Resende, 2007; Resende & Duarte, 2007; Piepho et al., 2008; Nunes et al., 2008; Kelly et al., 2009; Felipe & Duarte, 2010; Pinheiro et al., 2013). Segundo Piepho et al. (2008), em culturas anuais a metodologia de modelos mistos (BLUP) não é tão importante porque, em geral, se produz grande quantidade de dados fenotípicos por genótipo, obtidos de uma série de experimentos, e, assim, os diferentes métodos de estimação nem sempre fornecem resultados muito discrepantes, o que dá a falsa impressão de que a escolha do estimador não seja importante. Duarte (2010) relata que dentre as vantagens da abordagem de modelos mistos, é necessário previamente caracterizar que, muitas vezes, o cenário da pesquisa em melhoramento de plantas não é compatível com a tradicional metodologia de modelos fixos.



### 2.2.1 Modelagem estatística em modelos lineares mistos

O modelo estatístico que descreve a equação de modelos lineares mistos normalmente é apresentado como:

$$y = X\beta + Z\gamma + \varepsilon$$

em que:

$y$ : é o vetor<sub>(nx1)</sub> de observações;

$\beta$ : é o vetor<sub>(px1)</sub> paramétrico de efeitos fixos, com matriz de incidência  $X$  de ordem  $n \times p$ , conhecida;

$\gamma$ : é o vetor<sub>(qx1)</sub> paramétrico de efeitos aleatórios, com matriz de incidência  $Z$  de ordem  $n \times q$ , conhecida;

$\varepsilon$ : é um vetor<sub>(nx1)</sub> aleatório de erros, com média zero e matriz de variâncias-covariâncias  $R = I\sigma_e^2$ , sendo  $\sigma_e^2$  a variância dos efeitos residuais e  $I$  uma matriz identidade;

Nessa modelagem ainda se assume, normalmente, que as esperanças das variáveis estimadas sejam: que  $E(y) = X\beta$ ,  $E(\gamma) = E(\varepsilon) = \phi$ ; que os erros são independentemente distribuídos com média zero e matriz de variância-covariância  $R$ , i.e.,  $\varepsilon \sim N(\phi, R)$ ; que os efeitos aleatórios  $\gamma$  também são distribuídos com média zero e matriz de variância-covariância  $G$ , i.e.,  $\gamma \sim N(\phi, G)$ ; e  $Var(y) = V = ZGZ' + R$ . Definem-se, portanto,  $G$ , a matriz de variância-covariância dos efeitos aleatórios;  $R$ , a matriz de variância-covariância dos erros; e,  $\phi$ , um vetor nulo. Noutra forma,  $G$  é a matriz de variâncias e covariâncias genéticas entre os níveis dos efeitos aleatórios e  $V$  a matriz de variâncias e covariâncias fenotípicas entre as observações.

Para solução do modelo e consequente estimação de  $\gamma$  e  $\beta$ , torna-se necessária a inversão da matriz  $V$ , o que, normalmente, em aplicações, é de difícil solução devido à ordem dessa matriz. Assim, uma alternativa foi proposta por Henderson para solução desse sistema, pelo uso das equações de modelos mistos (EMM). Logo, assumindo-se  $G$  e  $R$  como conhecidas, a simultânea estimação de efeitos fixos (BLUE) e a predição dos efeitos aleatórios (BLUP) podem ser obtidas por meio da solução de EMM (Henderson, 1984):

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Com base nas EMM, os melhores preditores lineares não viciados (BLUP) de  $\gamma$ , que representam a solução dos efeitos aleatórios, são obtidos pela expressão  $BLUP(\gamma) = \hat{\gamma} = GZV^{-1}(y - X\hat{\beta})$ ; ou  $BLUP(\gamma) = \hat{\gamma} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{\beta})$ , neste caso, sem o uso da matriz  $V$ . Enquanto a obtenção dos melhores estimadores lineares não viciados (BLUE) associados ao vetor  $\beta$ , de efeitos fixos, resulta da solução de EMM por quadrados mínimos generalizados:  $\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}y$ ; ou, de outro modo,  $\hat{\beta} = [X'R^{-1}X - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}]^{-1}[X'R^{-1}y - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}y]$  (Mrode, 1996; Searle et al., 1992); em que  $(X'V^{-1}X)^{-}$  é uma matriz inversa generalizada de  $(X'V^{-1}X)$  (Searle, 1982).

Assumindo-se que os erros sejam não correlacionados, ou seja,  $R = I\sigma^2$ , as equações de modelos mistos passam a depender exclusivamente da matriz de variâncias e covariâncias genéticas  $G = A\sigma_g^2$  e de sua inversa  $G^{-1}$ . Logo, o sistema de equações pode ser reduzido a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix},$$

ou, ainda, considerando o componente de variância genética aditiva ( $\sigma_a^2$ ) ou o coeficiente de herdabilidade ( $h^2$ ) (Mrode, 1996):

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1} \frac{\sigma^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1} \frac{(1-h^2)}{h^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}.$$

### 2.2.2 Matrizes $G$ e $R$ em modelos lineares mistos

Para aplicação da abordagem de modelos mistos é necessário conhecer os componentes de variância associados às matrizes  $G$  e  $R$ . Assim, o primeiro passo para a sua aplicação normalmente é a obtenção de estimativas de  $G$  e  $R$  (Searle et al., 1992; André, 1999; Duarte, 2000; Resende, 2002, 2007). Em geral, os métodos de estimação dos componentes de  $G$  e  $R$  determinam o sucesso da predição, pois a eficiência de preditores (BLUP), para fins de seleção, depende de boas estimativas dos componentes de variância. Piepho et al. (2008) relatam que, com respeito aos componentes de variância, essenciais para a obtenção de predições seguras, tem-se que, em plantas, a avaliação é de menor

precisão devido ao reduzido número de genótipos avaliados, de mesma origem (população), e à estrutura de covariâncias serem mais complexas. Já no melhoramento de animais, por exemplo, avalia-se grande número de genótipos provenientes de uma única origem (progenitor) e, assim, os componentes de variância tornam-se bastante precisos, aproximando-se dos valores paramétricos.

Existem vários métodos de estimação de componentes de variância (métodos dos momentos, MINQUE, MIVQUE, ML, REML, VEIL), sendo suas propriedades, vantagens e desvantagens comentadas por André (1999), Duarte (2000) e Camarinha Filho (2002). Apesar dessa variedade de métodos, tem-se demonstrado maior interesse e aplicação ao método de máxima verossimilhança restrita (REML) (Patterson & Thompson, 1971), associado ao algoritmo *expectation-maximization* (EM) (Dempster et al., 1977; Bernardo, 2002). Este algoritmo é executado num processo iterativo que usa as estimativas iniciais de variâncias (genética e residual) para re-estimar os efeitos de  $\gamma$  e  $\beta$ ; os quais são usados em novo ciclo iterativo, que resulta em novas estimativas dos componentes de variância, até que a convergência do processo seja alcançada (ex. diferenças menores que 0,001). Com isso, obtém-se o melhor ajuste das estimativas dos componentes de variâncias. Além de EM, o processo iterativo pode ser executado com aplicação de outros algoritmos, por exemplo, *Fisher-scoring* (FS), Newton-Raphson (NR), ou informação média (AI, *average information*) (Smith et al., 2005; Resende, 2007; Butler et al., 2009).

Nessa condição, Piepho et al. (2008) relatam que, na prática, os termos BLUE e BLUP precisam ser substituídos por BLUE e BLUP empíricos, respectivamente, uma vez que os componentes de variância de  $G$  e  $R$  são desconhecidos e precisam ser substituídos por suas estimativas; preferencialmente obtidas por REML. Littell et al. (1996) indicam que nesse caso é apropriado substituí-los por *EBLUE* (*empirical best linear unbiased estimator*) e *EBLUP* (*empirical best linear unbiased predictor*), respectivamente. O termo empírico é adicionado com o intuito de indicar esse tipo de aproximação, que, na prática, normalmente é ignorado ou contornado apenas com uma interpretação conservativa (Duarte, 2000).

Considerando-se a matriz  $G$  dada por  $G = A\sigma_g^2$ , na sua definição, quando não se dispõe de informação de relacionamento entre os tratamentos genéticos, considera-se a sua obtenção pelo ajuste de  $\hat{G} = I\hat{\sigma}_g^2$ , sendo  $I$  uma matriz identidade de ponderação da variância genética. Quando se dispõe da informação de parentesco ou de similaridade

genética, é conveniente ponderar  $\sigma_g^2$  pela matriz  $A$ , de forma a obter  $\hat{G} = A\hat{\sigma}_g^2$ . Nesta condição a matriz  $A$  é uma matriz de parentesco, que indica o grau de associação genética entre os indivíduos (ex. genótipos) associados ao vetor  $\gamma$  (André, 1999; Bernardo, 2002).

Na abordagem clássica de modelos mistos, Henderson (1984) definiu a matriz  $A$  pelos coeficientes de parentesco entre os genótipos  $x$  e  $y$ , obtidos diretamente da informação de genealogias dos indivíduos. Conforme Silva Filho (2004), a matriz de similaridade genética entre indivíduos pode ser construída com base no coeficiente de parentesco de Malécot ( $r$ ), desde que informações de genealogia estejam disponíveis. Também pode ser usado o coeficiente de parentesco de Wright ( $f_{xy}$ ). Este coeficiente

equivale a  $f_{xy} = \frac{2r_{xy}}{\sqrt{(1+F_x)(1+F_y)}}$ , e consiste na probabilidade de que dois alelos num

mesmo loco sejam idênticos por descendência (Falconer & Mackay, 1996), o que reflete a covariância entre parentes (Kempthorne, 1957; Bernardo, 1994; Miranda et al., 2007; Nunes et al., 2008; Piepho et al., 2008). Piepho et al. (2008) sugerem que nem sempre os genótipos (tratamentos genéticos) atendem às suposições impostas à estimativa do coeficiente de parentesco; portanto, nessas condições podem-se obter estimativas de  $f_{xy}$  aproximadas, o que gerará previsões BLUP intermediárias àquelas esperadas numa condição em que as informações sejam fidedignas e adequadas.

Visto que nem sempre informações de genealogia estão disponíveis, esta era uma das limitações para uso dessa abordagem. Contudo, com o advento dos marcadores moleculares, possibilitou-se o uso de informação de coeficiente de parentesco obtida a partir desses marcadores (Bernardo, 1994; Ritland, 1996); ou de dissimilaridade genética (André, 1999; Silva Filho, 2004; Carneiro et al., 2006) para definição de  $A$ . Nessa condição a matriz  $A$  é substituída por uma matriz que reflete o coeficiente de parentesco ou a similaridade alélica entre indivíduos, considerando a identidade alélica por estado (André, 1999; Carneiro et al., 2006). Segundo Piepho et al. (2008), o uso de marcadores moleculares na obtenção da matriz  $A$  é potencialmente superior ao uso da genealogia, pois permite obter a contribuição genômica atual de um dado genitor para cada indivíduo relacionado. Bernardo (1994) destacou que, com o uso de marcadores moleculares, pode-se obter valores mais adequados do coeficiente de parentesco, em situações em que efeitos de seleção ou deriva genética ocorreram no desenvolvimento do genótipo.

Dentre as vantagens do uso da matriz de parentesco  $A$  obtida por marcadores moleculares ainda se acrescentam (André, 1999; Carneiro et al., 2006):

- a) permite o uso de toda a informação de semelhança alélica, convertida em identidade por descendência, existente entre os indivíduos avaliados;
- b) favorável quando não se dispõe de informações de genealogia dos indivíduos avaliados;
- c) evita a obtenção de estimativas ou predições inadequadas devido ao uso de informações erradas de genealogia.

Carneiro et al. (2006) relatam que erros de genealogia na definição da matriz  $A$  prejudicam a estimação e, consequentemente, geram prejuízos na seleção baseada nessas estimativas. Nessa condição, esses autores reforçam que o uso da matriz  $A$  obtida com informação advinda de marcadores moleculares representa uma estimativa mais realista da porcentagem de alelos em comum entre os indivíduos.

Dentre as vantagens do uso da abordagem de modelos mistos, em relação ao modelo clássico de análise de variância (ANOVA), encontra-se, ainda, a possibilidade de adoção de estruturas que considerem a existência de erros correlacionados, advindos, por exemplo, da presença de autocorrelação espacial nos dados. A autocorrelação espacial é caracterizada pela similaridade ou correlação entre observações de parcelas vizinhas, culminando na não independência entre observações, e, consequentemente, dos erros. Logo, tal como numa modelagem por quadrados mínimos ponderados ou generalizados, mesmo em modelo fixo, a adoção de modelos mistos permite acomodar diferentes estruturas para a matriz de variância-covariâncias dos erros ( $R$ ). As estruturas de  $R$  podem ser estimadas ajustando-se funções de semivariância ou de autocorrelação, por métodos de análise espacial (ex. técnicas geoestatísticas) ou de séries temporais. Assim viabiliza-se a análise dos dados mesmo numa condição de ausência de independência dos erros (Duarte, 2000; Resende, 2002; Resende, 2007). Alguns autores consideram essa suposição como necessária no cenário convencional de análise de variância por quadrados mínimos (Searle et al., 1992; Ramalho et al., 2000; Resende, 2007).

A avaliação de modelos que flexibilizam a matriz  $R$ , com aplicação de técnicas de análise espacial aplicadas aos próprios dados experimentais, representa uma importante forma de se obterem estimativas ou predições acuradas, mesmo numa condição de erros correlacionados. Com isso, caracteriza-se uma importante forma de controle adicional da variação local, mediante análise, que é flexibilizada também no contexto de modelos

mistos (Duarte, 2000; Duarte & Vencovsky, 2005; Piepho & Williams, 2010). Justifica-se esta aplicação pelo fato de que, nem sempre, a casualização e as formas de controle local baseadas em blocagem são suficientes para a neutralização dos efeitos de correlação espacial, e, por conseguinte, para a geração de uma análise de variância fidedigna (Resende, 2007).

As variáveis com dependência espacial, denominadas variáveis regionalizadas, apresentam uma aparente continuidade no espaço, denotada pela tendência de a variável apresentar valores muito próximos em dois pontos vizinhos e muito diferentes em pontos distantes (Resende, 2007). No estudo da dependência espacial dessas variáveis, aplicam-se métodos de análises de séries temporais e métodos geoestatísticos. Dentre os métodos para análise espacial por séries temporais destacam-se os modelos autorregressivos integrados de médias móveis (ARIMA) e o autorregressivo de primeira ordem (AR1). Quanto aos métodos geoestatísticos, o modelo geométrico exponencial é tido como padrão para o ajuste de uma função de autocorrelação ao variograma experimental em ensaio de campo. Detalhes da aplicação desses métodos são descritos por Duarte (2000), Resende (2002), Resende (2004), Resende (2007) e Butler et al. (2009).

Na condição em que os erros não são independentes e, portanto, a estrutura diagonal  $R = I\sigma^2$  não é adequada na estimação ou predição de valores genotípicos, adota-se a estrutura  $R = \Sigma$ , em que  $\Sigma$  representa a estrutura que determina a dependência espacial. Consequentemente,  $\Sigma$  acomoda a estrutura de erros correlacionados e sua respectiva função de autocovariância ou autocorrelação utilizada para descrever a dependência espacial. Essa estrutura será, portanto, considerada conjuntamente no processo de estimação ou predição dos valores genotípicos. Assim, viabiliza-se a análise dos dados, acomodando o controle de variações não controladas por “blocagem” no campo experimental e, consequentemente, proporcionando melhor aproveitamento da informação disponível aos melhoristas.

### **2.2.3 Propriedades e vantagens da metodologia de predição REML/BLUP**

Bernardo (1994) indica que REML é o procedimento mais adequado para estimar componentes de variância genética, considerando efeitos de parentesco entre genótipos; pois, nessa condição assume-se que os efeitos são correlacionados e, portanto, não independentes. Resende (2002, 2004, 2007) enumera uma série de vantagens práticas

de se usar o método REML associado aos modelos mistos. A seguir destacamos algumas destas vantagens inerentes exclusivamente a este procedimento de estimação:

- a) permite a simultânea correção para os efeitos ambientais, estimação de componentes de variância e a predição de valores genéticos;
- b) permite lidar com estruturas complexas de dados (medidas repetidas em diferentes anos, locais e até delineamentos);
- c) pode ser aplicado a dados desbalanceados e a delineamentos não ortogonais;
- d) permite flexibilidade na modelagem, contemplando a análise de dados correlacionados, devido ao parentesco, a distribuição temporal ou espacial;
- e) permite ajustar modelos que não podem ser acomodados pela tradicional análise de variância (ANOVA).

Conforme indicado por Mrode (1996), os valores genéticos preditos,  $\hat{\gamma}$ , com a metodologia BLUP apresentam as seguintes propriedades gerais: a) *Best* – em que as médias maximizam a correlação entre os valores verdadeiros ( $\gamma$ ) e preditos ( $\hat{\gamma}$ ), ou ainda, minimizam a variância do erro das predições; b) *Linear* – os preditores são funções lineares das observações; c) *Unbiased* – valores preditos para uma variável aleatória ou estimativas de efeitos fixos são não viciados; d) *Prediction* – envolve a predição do valor genético verdadeiro. Segundo White & Hodge (1989), existem várias propriedades úteis do BLUP, referentes ao melhoramento genético vegetal, dentre as quais se destacam:

- a) permite a estimação e predição não viciadas em um procedimento único;
- b) poder levar em conta efeitos de seleção e endogamia ao longo das gerações, desde que o grau de relacionamento genético (informação de parentesco) entre os indivíduos avaliados seja conhecido;
- c) a correlação entre os valores genéticos preditos e os verdadeiros é máxima dentre as classes de preditores lineares não viciados;
- d) permite prever o valor genético de indivíduos observados ou não;
- e) as variâncias e os erros das predições BLUP são menores em relação a outros métodos de estimação;
- f) possui o menor erro quadrático médio dentre todos os preditores lineares não viciados;

- g) em dados multivariados, com distribuição normal, a probabilidade de selecionar o melhor indivíduo é máxima entre os estimadores lineares não viciados com média zero.

Dentre as vantagens de uso do método REML/BLUP no melhoramento de plantas, tem merecido destaque a característica de que este possui máxima acurácia seletiva, quando comparado a outros métodos. Além disso, apresenta a possibilidade de uso simultâneo de várias fontes de informação, por exemplo, experimentos conduzidos em um ou vários locais, em uma ou várias safras. Apresenta, ainda, a flexibilidade de utilizar, num só modelo, todas as informações disponíveis, permitindo acumular informações de dados desbalanceados, comuns nas diferentes etapas de programas de melhoramento genético (Bernardo, 2002; Resende, 2007).

Duarte (2010) destaca que, no campo das vantagens da abordagem de modelos mistos, é necessário previamente caracterizar que, muitas vezes, o cenário da pesquisa em melhoramento de plantas não é compatível com a tradicional metodologia de modelos fixos. Por exemplo, nas fases preliminares do processo de avaliação genotípica, em geral, tem-se: muitos tratamentos genéticos a serem testados; mistura de diferentes tipos de tratamentos; tratamentos genéticos relacionados por origem; pouco material de propagação disponível para teste; parcelas pequenas; uso de tratamentos não repetidos; delineamentos não-ortogonais e, ou, desbalanceados; além de perda de parcelas, entre outros. Piepho et al. (2008) relatam que, mesmo em etapas de avaliação final nos programas de melhoramento, com o uso de delineamentos ortogonais, os experimentos são caracterizados como notoriamente desbalanceados, devido ao descarte anual de genótipos com baixo desempenho; enquanto novas entradas são adicionadas aos experimentos para nova avaliação. Logo, numa rede de ensaios envolvendo locais e, ou, anos (ou épocas), o desbalanceamento e a falta de ortogonalidade são comuns. Duarte (2010) considera que nessas condições, a possibilidade de melhorar a qualidade das estimativas e previsões genotípicas, por meio do uso de informações interblocos e, ou, intertratamentos, passíveis de recuperação na abordagem de modelos mistos, é real e não deve ser desperdiçada.

Adicionalmente, possibilita o aproveitamento da informação de parentesco entre os genótipos avaliados, maximizando, em geral, o uso dos dados disponíveis no processo de seleção (Bernardo, 2002; Resende, 2007). Segundo André (1999), o BLUP considera a informação da covariância genética entre genitores (genótipos) que participam do processo seletivo; informações estas desprezadas por outras metodologias.



Em resumo, recentemente o procedimento BLUP tem tido maior aceitação no melhoramento de plantas, devido ao fato de que, sob condições de desbalanceamento, este preditor tem a vantagem de resultar em predições de maior acurácia do que simples estimativas obtidas por quadrados mínimos ordinários. Além disso, permite a incorporação de informações de parentesco, otimizando o uso dos dados (Bernardo, 2002).

Henderson et al. (1959), no entanto, alertam que, em caso de experimentos balanceados e não havendo ou não se conhecendo o parentesco entre os genótipos, as soluções do modelo misto (BLUP) e as soluções dos quadrados mínimos (BLUE) podem levar ao mesmo ordenamento dos genótipos. Conforme relatado por Silva Filho (2004), obtêm-se estimativas equivalentes para estimadores de quadrados mínimos, quadrados mínimos generalizados ou mesmo máxima verossimilhança, numa condição em que os erros são normalmente distribuídos e não ocorra estrutura de inter-relação (correlação) entre os tratamentos.

Apesar das inúmeras vantagens relacionadas à aplicação dos modelos mistos na análise de dados experimentais, alguns problemas potenciais também têm sido levantados. Duarte (2010) destaca o fato de que, em aplicações no melhoramento de plantas, normalmente as variâncias apresentam estimativas com intervalos de confiança amplos, além de serem obtidas diretamente dos dados, resultando apenas em predições empíricas (EBLUPs). Essa condição pode induzir a vícios na predição dos valores genotípicos e, conseqüentemente, na tomada de decisão acerca da seleção de genótipos. Adicionalmente, em estudos envolvendo várias populações, o efeito “shrinkage” associado às predições BLUP pode conduzir à seleção prioritária de genótipos aparentados, advindos principalmente das populações (origem) com médias superiores. Com isso, em longo prazo, pode haver diminuição da variabilidade genética, o que inspira maior atenção com respeito ao controle do tamanho efetivo populacional.

Com o advento da genotipagem em larga escala, por exemplo, com uso de *single nucleotide polymorphism* (SNP), Meuwissen et al. (2001) desenvolveram a abordagem da predição de valor genômico, utilizando-se seleção genômica ampla no contexto de modelos mistos. Estes autores também obtiveram a predição dos efeitos, BLUP de  $\gamma$ , pela solução das equações de modelos mistos (Henderson, 1984). De forma geral, BLUP permite a predição do efeito alélico mesmo quando se têm mais efeitos que dados amostrais ( $n$ ), sendo que, atualmente, essa condição já é comum devido à grande potencialidade de genotipagem, com altíssimos números de marcadores SNP.

Nesse contexto, consideram-se todas as marcas disponíveis para a estimação do efeito genético, sendo atribuído um valor comum à variância de cada alelo ( $V_g/n$ ); isto é dado pela relação entre variância genética total ( $V_g$ ) e o número de marcas ( $n$ ). Assim, o uso de BLUP permite a obtenção de predições com boa acurácia, sendo indicado para aplicações em predição de valores genéticos de caracteres quantitativos que apresentam baixa herdabilidade (Meuwissen et al., 2001; Bernardo & Yu, 2007; Lorenzana & Bernardo, 2009). Meuwissen et al. (2001) destacaram que o uso de BLUP não possibilita a identificação da posição de QTL. Esses autores ressaltaram, ainda, que a posição correta do QTL não é essencial quando se busca acessar o seu valor, permitindo, assim, a obtenção de predições de boa acurácia via BLUP.

#### 2.2.4 Critérios de escolha e seleção de modelos de análise

Dentre as vantagens da aplicação de modelos mistos encontra-se a flexibilidade do uso de estruturas de variâncias e covariâncias dos efeitos aleatórios, por meio das matrizes  $G$  e  $R$ , em diversas modelagens. Assim, permite-se obter diferentes soluções num mesmo conjunto de dados. Portanto, a definição de critérios de escolha de qual dos possíveis modelos que melhor se ajustem à situação avaliada torna-se de grande importância para uma correta análise e tomada de decisão. Em geral, a escolha do modelo com melhor ajuste aos respectivos conjuntos de dados, nas diferentes configurações de estruturas de  $G$  e  $R$ , pode ser feita com base no teste da razão de verossimilhanças ( $LRT$ , *likelihood ratio test*) e, ou, nos critérios de informação de Akaike ( $AIC$ ) (Akaike, 1974) ou Bayesiano ( $BIC$ ) (Schwarz, 1978).

O  $LRT$  compara estimativas obtidas por máxima verossimilhança, para diferentes modelos, sendo um dos modelos, uma versão restrita do outro. De acordo com Resende (2007), esse teste permite comparar modelos que tenham uma estrutura hierárquica em relação aos fatores aleatórios e fixos. Nessa condição, um modelo tem  $r$  parâmetros adicionais, e o teste verifica se esses parâmetros adicionais melhoram significativamente o modelo (Camarinha Filho, 2002). Assim, a comparação entre os modelos é realizada com aplicação do  $LRT$ , que é obtido por:  $LRT = 2\log(\ell_{R2}/\ell_{R1}) = 2[\log(\ell_{R2}) - \log(\ell_{R1})]$ , em que  $\ell_{R1}$  representa a função de máxima verossimilhança do modelo aninhado (ou restrito); e,  $\ell_{R2}$ , a função de máxima verossimilhança do modelo de referência. Nessa função, a estatística  $LRT$  calculada

apresenta assintoticamente distribuição de  $\chi^2$ , em que o número de graus de liberdade é obtido subtraindo-se o número de parâmetros do modelo aninhado daquele do modelo de referência. Com isso, é possível verificar a hipótese de nulidade entre as capacidades de ajuste dos modelos, por meio da respectiva probabilidade ( $p$ -valor) associada ao teste (Butler et al., 2009).

O critério *AIC* é obtido por:  $AIC = -2\ell_{Ri} + 2t_i$ ; e o critério *BIC* por:  $BIC = -2\ell_{Ri} + t_i \log(n - p)$ , em que  $t_i$  representa o número total de parâmetros de efeitos fixos e aleatórios no modelo;  $n$ , o número de observações; e  $p$ , o posto da matriz  $X$ . Esses critérios são obtidos pela maximização dos critérios de informação de máxima verossimilhança restrita, dependendo também dos números de observações e de parâmetros do modelo. Assim, é escolhido como adequado o modelo que apresenta o menor valor para tais critérios (Butler et al., 2009). Esses critérios estão fundamentados na teoria de decisão e desfavorece modelos com número grande de parâmetros, para se evitar excesso de parametrização, e favorece a escolha de modelos parcimoniosos (Camarinha Filho, 2002). Segundo Duarte & Vencovsky (2005), modelos sobreparametrizados apenas se justificam caso demonstrem uma melhoria significativamente importante na qualidade de ajustamento, em relação a modelos mais simples. Assim, indicam a adoção dos critérios comparativos *AIC* e *BIC*, que tendem a penalizar modelos mais parametrizados.

Resende (2007) indica que os critérios *LRT* e *AIC* são complementares por suas características. Por exemplo, *LRT* exige modelos hierarquizados (aninhados), em que um modelo é caso especial ou particular do outro, permitindo, assim, a aplicação de teste de probabilidade e consequente definição dos níveis de significância associados. Já o critério *AIC* não exige hierarquização, mas também não permite testes de probabilidade. Piepho & Williams (2010) relatam que, tanto *LRT*, como *AIC*, são procedimentos padrões para avaliar a relevância de modelos. Esses autores indicam que a minimização do critério *AIC* é uma estratégia que permite identificar um modelo aproximado com menor discrepância de um provável modelo real, sendo bastante parcimonioso para ser suportado pelos dados.

### 3 DIVERSIDADE E ESTRUTURA GENÉTICA EM GERMOPLASMA DE SOJA

#### RESUMO

O conhecimento da variabilidade genética disponível é de grande importância para o planejamento de hibridações num programa de melhoramento. Nesse contexto, o objetivo deste estudo foi avaliar, por meio de marcadores microssatélites (SSR), a diversidade genética em germoplasma de soja adaptado às condições brasileiras, com vista ao dimensionamento de seu potencial de uso. O material experimental constituiu-se de 192 acessos: 26 cultivares, incluindo introduções (grupo A) e cultivares brasileiras (grupo B); e 166 linhagens endogâmicas, que se originaram de cruzamentos entre genitores introduzidos (grupo AA), entre genitores brasileiros (grupo BB) e entre genitores introduzidos e brasileiros (grupo AB). A genotipagem foi realizada com uso de 37 marcadores (locos) SSR. Para avaliação da divergência genética, considerou-se a análise descritiva dos locos. Também estimou-se a distância genética de Rogers, modificada por Wright, obtendo-se dendrograma pelo método de agrupamento e critério UPGMA. Nos 37 locos SSR foram identificados 222 alelos, com média de seis alelos por loco e variação de 2 a 14 alelos. Desses alelos, 115 apresentaram frequências abaixo de 0,10, dos quais 28 com frequência inferior a 0,01; a maioria em acessos introduzidos. O agrupamento dos genótipos mostrou-se associado à origem do germoplasma, resultando em dois grupos: i) A e AA; e ii) B, AB e BB. Este resultado foi concordante com a análise de estruturação genética dos acessos, os quais se estruturaram em duas subpopulações ( $K=2$ ). A diversidade genética média entre genótipos foi de 0,77. Das estimativas de distâncias genéticas, 85% foram superiores a 0,70, indicando bom potencial do germoplasma para hibridações em programas de melhoramento genético da soja nessa região. Concluiu-se ainda que os marcadores SSR são úteis na identificação de grupos genotípicos divergentes, bem como de combinações de alta variabilidade genética. Ademais, demonstra-se que o uso de germoplasma introduzido garante a incorporação de alelos necessários à ampliação da base genética da espécie e, conseqüentemente, da variabilidade necessária para uso no processo seletivo.

*Palavras-chave:* variabilidade genética, *Glycine max*, microssatélites, introdução de germoplasma, base genética.

#### ABSTRACT

#### DIVERSITY AND STRUCTURE GENETIC IN SOYBEAN GERMPLASM

Knowledge of the available genetic variability is very important for planning hybridizations in a breeding program. In this context, microsatellite markers (SSR) were used in the present study to assess the genetic diversity among soybean germplasm adapted

to Brazilian conditions for potential use in a crossing program. The experimental material consisted of 192 accessions, which included 26 cultivars among introductions (group A) and Brazilian cultivars (group B); and 166 inbred lines derived from crosses among introduced parents (group AA), Brazilian parents (group BB) and introduced versus Brazilian parents (group AB). Thirty-seven SSR markers (loci) were used in genotyping the 192 accessions. The genetic divergence were assessed using the descriptive analysis of the loci. The Rogers genetic distance modified by Wright was also estimated to build a dendrogram in the cluster analysis method using UPGMA criteria. A total of 222 alleles were identified in the 37 SSR loci, with a mean of six alleles and a range of 2 to 14 alleles per loci. Of these 222 alleles, 115 presented frequencies below 0.10 of which 28 had frequency below 0.01, mostly in introduced accessions. The genotype clustering showed association to the origin of the germplasm, and resulted in two groups: i) A and AA; and ii) B, AB and BB. This result was in agreement with the analysis of genetic structure of the accessions, that grouped them into two subpopulations ( $K=2$ ). The mean genetic diversity among genotypes was 0.77. Eighty five percent of the genetic distances estimates were greater than 0.70, indicating a good potential of this germplasm for hybridization in soybean breeding programs of Central Brazil. In addition, it was concluded that SSR markers are useful in clustering genotypes in divergent groups, as well as identifying genotype combinations with high genetic variability. Furthermore, it was shown that the use of introduced germplasm guarantees incorporation of alleles needed to widen the genetic base of the species, and consequently the variability necessary for use in the selective process.

**Key words:** genetic variability, *Glycine max*, microsatellites, germplasm introduction, genetic base.

### 3.1 INTRODUÇÃO

Apesar da grande importância econômica da soja no mundo, esta espécie cultivada é comumente reportada na literatura como sendo de base genética estreita. Wysmierschi (2010), ao comparar a base genética e o coeficiente médio de parentesco da soja em alguns países (Estados Unidos, China, Japão, Índia e Brasil), relata que a China representa a maior base genética, enquanto o Brasil, a base genética mais estreita.

Dentre as razões para o estreitamento dessa base genética, Hyten et al. (2007) destacam o baixo nível de variabilidade genética na própria espécie selvagem, *Glycine soja*, se comparada a outras espécies vegetais. Acresce-se a isto o afunilamento da diversidade genética (*bottleneck*) ocorrido durante a sua domesticação, e, recentemente, devido ao processo de melhoramento genético (Hyten et al., 2007; Guo et al., 2010). O melhoramento tem contribuído para esse efeito, sobretudo pelo pequeno número de genótipos ancestrais (*landraces* ou *plant introductions* – *PI*) que foram introduzidos dos centros de origem da espécie, para a formação da base inicial dos principais programas de

melhoramento no mundo. Atualmente, esses programas usam como genitores apenas um grupo limitado de genótipos elite, basicamente cultivares de alta “performance” (Hiromoto & Vello, 1986; Gizlice et al., 1994; Hyten et al., 2007; Mikel et al., 2010; Wysmierschi, 2010).

Com a domesticação ocorreu a maior perda de diversidade genética na soja (Hyten et al., 2006; Guo et al., 2010). Esses estudos indicam redução de 50% da diversidade da espécie e perda de 81% de alelos raros (com frequência inferior a 0,05). Guo et al. (2010) verificaram que a diversidade genética em *landraces* foi equivalente à metade da encontrada na espécie selvagem, sugerindo forte afunilamento da diversidade. Também observaram drástica redução na heterozigosidade esperada, tendo detectado nas *landraces* estudadas, apenas 41,9% da riqueza alélica da espécie selvagem. Hyten et al. (2006) avaliaram a diversidade genética em genótipos selvagens, *landraces* e cultivares melhoradas americanas (obsoletas e modernas), tendo constatado que as cultivares modernas retiveram 72% da diversidade presente em *landraces* de origem asiáticas; contudo, houve perda de 79% de alelos raros (neste caso, considerados como de frequência inferior a 0,10). Eles relacionaram a perda de diversidade ao pequeno número de introduções asiáticas usadas na formação inicial dos programas de melhoramento dos Estados Unidos, e não à seleção artificial imposta pelo processo de melhoramento. Isso porque não detectaram diferença expressiva entre as diversidades de cultivares ancestrais (antigas e obsoletas) e cultivares elite (modernas).

Gizlice et al. (1994) observaram que cerca de 75% dos genes das cultivares modernas, nos Estados Unidos, advêm das dezessete primeiras linhagens desenvolvidas no país, antes de 1960. Caracterizaram, ainda, que 95% dos genes encontrados em cultivares modernas foram provenientes de 35 genótipos, dos quais 28 são ancestrais (*PI* ou *landraces*), e sete, provenientes de melhoramento. Nesse grupo, quatro cultivares contribuíram com cerca de 40% dos genes das cultivares norte americanas modernas. Também nos EUA, Mikel et al. (2010) verificaram que, entre os anos de 1970 e 2008, dez cultivares representaram 16,8% dos genitores usados no desenvolvimento de novas cultivares. No período de 1999 a 2008, constataram que 23% destas tiveram, no mínimo, 25% de contribuição genética de um único genitor (cultivar A3127).

A base genética da soja brasileira tem origem principal em cultivares introduzidas dos Estados Unidos (Hiromoto & Vello, 1986; Hyamanaka et al., 2007). Segundo esses autores, as cultivares brasileiras foram derivadas, basicamente, de 26

introduções de origem norte americana. Destas, apenas quatro representavam aproximadamente 50% do *pool* gênico brasileiro. Acrescentando-se mais cinco, concluíram que nove delas foram responsáveis por mais de 80% do conjunto de genótipos amostrado. Em estudo mais recente, Wysmierski (2010) verificou a presença de sessenta ancestrais comuns na genealogia de 444 cultivares brasileiras. Destas, quatorze linhagens ancestrais representam 92,41% da base genética brasileira, e, apenas quatro ancestrais contribuíram com 55,26%.

Considerando-se que o *pool* gênico da soja brasileira é derivado principalmente de cultivares norte americanas, é natural, portanto, que compartilhe também de sua baixa diversidade. Vello et al. (1988) constataram que a base genética de cultivares brasileiras era representada por uma população com tamanho efetivo ( $N_e$ ) entre onze e quinze; isto é, relativamente baixo. Posteriormente, Miranda (2005) obteve estimativa de  $N_e$  entre treze e dezesseis, justificando o pequeno aumento pela provável incorporação de novos ancestrais na base genética brasileira. Contudo, em estudo posterior, Miranda et al. (2007) obtiveram novamente estimativas de  $N_e$  entre onze e treze. Assim, concluíram pela existência de alto grau de similaridade genética entre as principais cultivares de soja recomendadas para cultivo no Brasil.

Apesar dos diversos relatos de reduzida base genética, os programas de melhoramento de soja têm obtido sucesso na seleção de cultivares com alto potencial produtivo. Nos Estados Unidos, o ganho genético médio foi de 3,2% por ciclo de melhoramento (Mikel et al., 2010). Na Argentina, este ganho, em produtividade de grãos, foi de 1,2% a 2,0% ao ano (Ferrarotti, 2004; Rossi, 2004). No Brasil, considerando-se apenas o Estado do Paraná, os ganhos genéticos entre 1981 e 1986 foram estimados em 1,8% e 1,3%, nos grupos de maturação precoce e tardio, respectivamente (Toledo et al., 1990). Em outro estudo, Toledo et al. (2004) concluíram que os programas brasileiros de melhoramento de soja foram eficientes no desenvolvimento de cultivares, tendo incrementado ganhos genéticos em produtividade de grãos na magnitude de 1,5% a 2,0% ao ano.

Dentre as estratégias para a avaliação da diversidade genética encontram-se aquelas com uso de marcadores moleculares, com destaque para os microssatélites (SSR). Estes marcadores têm sido amplamente utilizados, sobretudo por sua elevada capacidade de detecção de polimorfismo e seu caráter multialélico (Ferreira & Grattapaglia, 1998). Em estudos com soja, Hwang et al. (2009) e Priolli et al. (2010) enumeram algumas

características desses marcadores, que favorecem o seu emprego na estimação da diversidade genética: abundância no genoma; alto nível de polimorfismo; codominância; detecção por PCR; posição conhecida no genoma; e capacidade de automatização da leitura. Tudo isso favorece a identificação fácil e precisa de diferentes alelos por loco. Ademais, são marcadores altamente efetivos na distinção de acessos com estreita base genética, como é o caso do germoplasma brasileiro de soja (Priolli et al., 2010).

Nesse contexto, o objetivo desta pesquisa foi avaliar, por meio de marcadores microsatélites (SSR), a diversidade e a estruturação genética em amostra de germoplasma de soja adaptado ao cultivo no Brasil, com vista ao dimensionamento do potencial da variabilidade genética disponível, visando o seu emprego em programas nacionais de melhoramento genético da cultura.

## 3.2 MATERIAL E MÉTODOS

### 3.2.1 Material experimental

O material experimental em avaliação constituiu-se de 192 genótipos. Estes estão divididos em: a) 26 cultivares, sendo doze introduções provenientes de regiões temperadas (grupo A) e quatorze cultivares comerciais brasileiras (grupo B); e b) 166 linhagens endogâmicas, em gerações  $F_8$  a  $F_{10}$ , das quais 35 são oriundas de cruzamentos exclusivamente realizados entre genitores introduzidos (grupo AA), 68 de cruzamentos entre genitores brasileiros (grupo BB) e 63 de cruzamentos entre genitores introduzidos e brasileiros (grupo AB). Todos os genótipos são portadores de transgenia para tolerância ao herbicida glifosato (tecnologia Roundup Ready ou RR) ou com esta característica, acrescida de resistência a insetos desfolhadores (tecnologia Intacta RR2 Pro).

As introduções são cultivares elite modernas, oriundas da Argentina (11) e Estados Unidos (1), e atualmente recomendadas para cultivo no Brasil. Juntamente com as cultivares brasileiras, representam as principais cultivares modernas atualmente cultivadas na região de Cerrado do Brasil Central. Dentre estas cultivares, onze representaram mais de 41% da área plantada com soja no país, na safra 2012/13 (Kleffmanngroup, 2013).

As linhagens endogâmicas são provenientes do programa de melhoramento genético de soja da empresa Nidera Sementes Ltda. Estas foram selecionadas com o objetivo de obtenção de novas cultivares adaptadas à referida região produtora de soja no Brasil.



### 3.2.2 Extração de DNA e genotipagem

O DNA genômico dos 192 genótipos foi extraído de folhas jovens (novas), com uso do método CTAB, conforme protocolo descrito por Doyle & Doyle (1990). Para isso, foram utilizados 100 mg de tecido foliar, que foi macerado com auxílio do equipamento TissueLyser (Applied Biosystems), após adição de 500  $\mu\text{L}$  de CTAB (concentração a 2%). Após a extração, o DNA foi quantificado em gel de agarose a 1% e, posteriormente, sua concentração foi ajustada para 10  $\text{ng} \cdot \mu\text{L}^{-1}$ .

A reação de PCR (reação de polimerase em cadeia) foi realizada adicionando-se cerca de 30  $\text{ng}$  do DNA; 0,16  $\mu\text{mol} \cdot \text{L}^{-1}$  do inicializador ou *primer forward*; 0,20  $\mu\text{mol} \cdot \text{L}^{-1}$  do *primer reverse*; 0,13  $\mu\text{mol} \cdot \text{L}^{-1}$  do *primer* M13 (IRDye 700 ou IRDye 800); 250  $\mu\text{mol} \cdot \text{L}^{-1}$  de cada desoxirribonucleosídeo trifosfato (dNTP); 0,2  $\mu\text{mol} \cdot \text{L}^{-1}$  de BSA, solução tampão de PCR contendo 50  $\text{mmol} \cdot \text{L}^{-1}$  de KCl, 10  $\text{mmol} \cdot \text{L}^{-1}$  de Tris-HCl (pH 8,9), 2,0  $\text{mmol} \cdot \text{L}^{-1}$  de  $\text{MgCl}_2$  e uma unidade da enzima Taq DNA-polimerase, em volume total de 20  $\mu\text{L}$ . O ciclo total de amplificação foi realizado em termociclador MJ96+ (Biocycler), programado para iniciar com cinco minutos a 94° C, seguidos por dez ciclos com um minuto de desnaturação a 94° C, um minuto de pareamento a 50° C (decréscimo de 1° C por ciclo) e um minuto de extensão a 72° C; acrescidos de trinta ciclos de quarenta segundos a 94° C, quarenta segundos a 40° C e um minuto a 72° C, além de extensão final de dez minutos a 72° C.

A genotipagem foi realizada com uso de 37 marcadores SSR (Satt545, Satt599, Satt187, Satt429, Satt197, Satt453, Satt577, Satt180, Satt294, Satt277, Satt281, Satt184, Satt129, Satt216, Satt600, Satt386, Satt389, Satt263, Satt146, Satt510, Satt012, Satt038, Satt192, Satt434, Satt354, Satt414, Sat\_224, Satt196, Satt588, Satt229, Satt462, Satt308, Satt463, Satt009, Satt022, Satt243, Satt345). Estes foram previamente selecionados por sua posição no genoma e pelo conteúdo de polimorfismo informativo (PIC) (Cregan et al., 1999; Hwang et al., 2009; Sayama et al., 2011). Inicialmente preconizou-se o uso daqueles com maior valor para diversidade genética (ou PIC) e, pelo menos, um em cada braço cromossômico. Isto resultaria em quarenta marcadores, embora apenas os 37 anteriormente citados tenham sido utilizados, pois três deles apresentaram baixa qualidade de genotipagem.

A genotipagem para os 37 marcadores foi realizada com uso de géis de poliacrilamida, contendo tampão TBE 1x e 7,0  $\text{mol} \cdot \text{L}^{-1}$  de uréia, com eletroforese em

equipamento 4300S DNA Analyser (Li-cor Corporate). Previamente, amostras de cada dois locos, representadas pelos dois tipos de marcação (IRDye 700 ou 800), foram preparadas em solução com 1,0 µL de cada produto da PCR e 3,0 µL de *Blue Stop Solution* (Li-cor Corporate). Em seguida, foram desnaturadas a 95° C, por três minutos. Posteriormente, em cada gel foram aplicadas 96 amostras e dois marcadores de peso molecular conhecido (50-350 bp IRDye700 e IRDye800, Li-cor Corporate). Por fim, com uso de programas computacionais específicos (SAGA Lite e SAGA MX Generation, Li-cor Corporate), efetuou-se a leitura automática dos locos, considerando-se fragmentos amplificados de diferentes tamanhos como sendo diferentes alelos de um loco simples.

### 3.2.3 Análise genética e populacional

A avaliação da divergência genética foi realizada mediante a análise descritiva dos 37 locos, estimando, para cada loco, o número de alelos com suas frequências, heterozigosidade, coeficiente de endogamia e índice de diversidade genética ( $h$ ) (Nei, 1973). As estimativas de  $h$ , para cada loco, foram obtidas por:  $h_{(i)} = 1 - \sum_j p_{ij}^2$ , em que  $p_{ij}^2$  é a frequência do  $j$ -ésimo alelo no  $i$ -ésimo loco, somada para todos os alelos nos respectivos locos (Nei, 1973; Weir, 1996).

A estimativa da diversidade genética ( $h$ ), que expressa uma medida equivalente ao conteúdo de polimorfismo informativo (PIC), é importante por fornecer o nível de polimorfismo potencial detectável pelo respectivo marcador. Essa estatística expressa, em uma única medida, o número de alelos por loco e suas frequências relativas, indicando o poder de discriminação de cada loco (Ferreira & Grattapaglia, 1998; Hwang et al., 2009). Geralmente, locos SSR com maiores valores de  $h$ , ou de PIC, possibilitam maior poder de discriminação de genótipos em análises genéticas; ou ainda, maior é a probabilidade de que exista polimorfismo entre quaisquer dois acessos para aquele loco.

A divergência genética entre pares de genótipos foi estimada pela distância

genética de Rogers, modificada por Wright (1978):  $d_w = \frac{1}{\sqrt{2m}} \sqrt{\sum_{i=1}^m \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2}$ , em

que,  $p_{ij}$  e  $q_{ij}$  representam as frequências alélicas de dois indivíduos relativas ao  $j$ -ésimo alelo do  $i$ -ésimo loco;  $n_i$  é o número de alelos no  $i$ -ésimo loco; e  $m$  representa o número de locos considerados na comparação entre pares de indivíduos. O estimador  $d_w$  apresenta

propriedades de distância euclidiana, sendo apropriado para estudos de divergência genética entre populações ou entre indivíduos dentro de populações, para os quais a informação alélica, neste caso advinda de marcador codominante (SSR), encontra-se disponível (Reif et al., 2005). Os valores de  $d_w$  variam entre zero e um, representando diretamente a distância genética entre pares de acessos. O valor zero indica genótipos idênticos e valores acima de zero, até um, indicam genótipos cada vez mais distintos.

A medida de divergência genética foi, inicialmente, determinada entre os cinco grupos de origem (A, B, AA, AB e BB), e, em seguida, entre os 192 genótipos, independentemente de sua distribuição nos grupos. Os respectivos dendrogramas foram obtidos pelo método de agrupamento SAHN (sequencial, aglomerativo, hierárquico e não sobreposto) e critério de ligação UPGMA (médias aritméticas não ponderadas). A qualidade dos respectivos agrupamentos foi avaliada por correlação cofenética, teste de Mantel (com 10 mil permutações) e consistência dos nós por *bootstrap* (com 10 mil reamostragens). Essas análises foram implementadas com uso dos aplicativos TFPGA (Miller, 2007), NTSYSpc (Rohlf, 2000) e Bood (Coelho, 2000).

A estruturação genética considerando os cinco grupos de origem (populações) foi avaliada pelo método de Pritchard et al. (2000). Este método permite identificar a ocorrência de estruturação genética entre populações, geralmente caracterizada pela presença de subpopulações, com frequências alélicas distintas, posicionando adequadamente os indivíduos às subpopulações, sem qualquer indicação prévia. Assim, obtém-se a estimativa de diversidade genética entre subpopulações ( $F_{ST}$ ), que fornece a ideia de ocorrência ou não de estruturação genética. Segundo Weir (1996), o parâmetro  $F_{ST}$  indica quanto da variação genética total é devida à diferenciação genética entre as subpopulações; logo, permite avaliar se há ou não estruturação genética entre as populações. Esta análise foi implementada no aplicativo Structure (Pritchard et al., 2000), simulando-se o número de populações ( $K$ ) entre um e cinco, com cinco iterações cada. O número  $K$  com melhor ajuste ao conjunto de dados foi definido pelo método de Evanno et al. (2005), que faz uso da estatística  $\Delta K$ . Esta permite identificar o número de classes ( $K$ ) com a maior verossimilhança associada, indicado pelo maior valor de  $\Delta K$ , o qual é tido como adequado para representar a subestruturação do conjunto de indivíduos (acessos).

### 3.3 RESULTADOS E DISCUSSÃO

#### 3.3.1 Análise genética descritiva e da diversidade genética

Os 37 locos microsatélites (SSR) usados na genotipagem dos acessos foram polimórficos. Nestes locos foram identificados 222 alelos, com média de seis alelos por loco. O número de alelos variou entre dois (Satt599 e Satt277) e quatorze (Satt129) (Tabela 3.1). Dentre os alelos, 115 apresentaram frequências abaixo de 0,10, sendo 28 destes com frequência inferior a 0,01. Apenas seis marcas, presentes nos locos Satt187, Satt196, Satt243, Satt277, Satt588, Satt229, tiveram frequências acima de 0,70. Não foi identificado qualquer alelo com frequência superior a 0,9, e, portanto, com chance de fixação nos grupos avaliados (Figura 3.1).

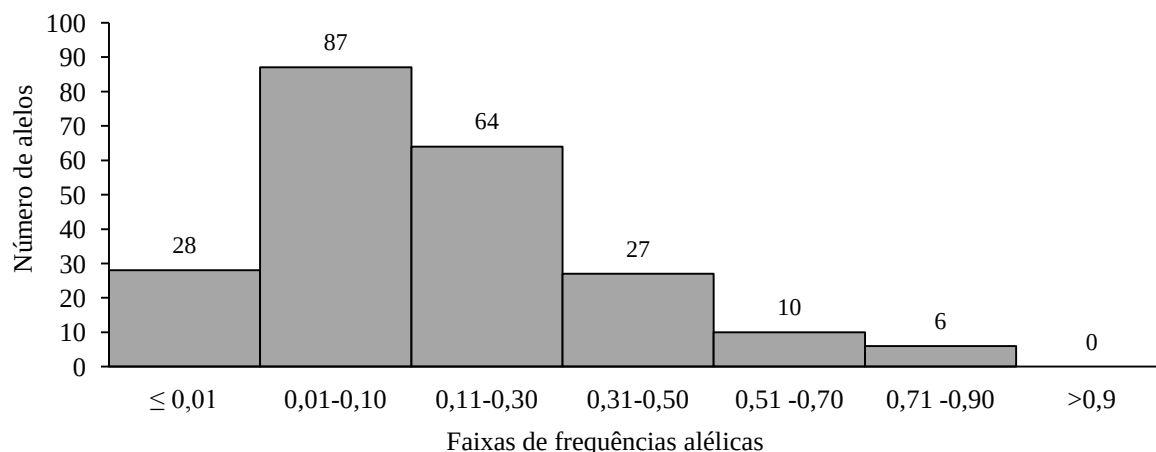
O número de alelos encontrado é concordante com aqueles identificados por outros autores, em estudos similares com genótipos melhorados. Yamanaka et al. (2007) avaliaram 272 genótipos, de origem da China, Japão ou Brasil, e identificaram entre três e treze alelos, com base em doze marcadores SSR. Priolli et al. (2010) obtiveram média de 5,1 alelos por loco, com variação entre dois e oito alelos, em avaliação de 168 cultivares brasileiras, com dezoito locos SSR. Garcia et al. (2007), avaliando germoplasma adaptado ao consumo humano, também obtiveram, em média, 5,1 alelos, em dez locos SSR previamente selecionados para maior informatividade. Mulato et al. (2010) identificaram entre dois e 21 alelos, com média de 8,63 alelos por loco, em introduções provenientes de diferentes regiões do mundo. Observaram, ainda, que houve significativo número de alelos raros ou exclusivos, associados a alguns dos acessos avaliados. Hwang et al. (2009), considerando 23 genótipos melhorados e um acesso selvagem, identificaram entre dois e sete alelos, nos 1.238 locos SSR avaliados, com média de 2,8 alelos por loco. Guo et al. (2010) obtiveram uma estimativa de riqueza alélica (alelos por loco polimórfico) de 11,8, para *landraces*, e 20,3 para acessos selvagens (*Glycine soja*). No total, identificaram, em média, 26,8 alelos por loco, numa faixa entre 5 e 56 alelos, nos 56 marcadores SSR usados.

Todos esses estudos caracterizam elevada variabilidade genética disponível nas diferentes fontes de germoplasma, incluindo *landraces*. Nesse sentido, Hyten et al. (2006) indicam o uso desse alto potencial de variabilidade proveniente de *landraces*, pois garante a incorporação de alelos relacionados a características específicas (resistência a doenças e nematoides) em germoplasma elite, principalmente pela disponibilização de alelos raros.

**Tabela 3.1.** Descrição dos grupos de ligação, número de alelos, índice de diversidade genética ( $h$ ), heterozigosidade observada ( $Het_o$ ), coeficiente de endogamia ( $f$ ) e faixas de tamanhos de alelos identificados com uso de 37 locos SSR, em 192 genótipos de soja.

|     | Loco <sup>1</sup> | Grupo de ligação <sup>1</sup> | Número de alelos | $h$  | $Het_o$ | $f$  | Faixa de tamanho |
|-----|-------------------|-------------------------------|------------------|------|---------|------|------------------|
| 1)  | Satt545           | A1                            | 7                | 0,66 | 0,03    | 0,96 | 196-220          |
| 2)  | Satt599           | A1                            | 2                | 0,47 | 0,01    | 0,98 | 191-203          |
| 3)  | Satt187           | A2                            | 3                | 0,42 | 0,00    | 1,00 | 262-313          |
| 4)  | Satt429           | A2                            | 5                | 0,70 | 0,02    | 0,98 | 259-286          |
| 5)  | Satt197           | B1                            | 8                | 0,74 | 0,04    | 0,94 | 130-217          |
| 6)  | Satt453           | B1                            | 4                | 0,56 | 0,02    | 0,96 | 251-272          |
| 7)  | Satt577           | B2                            | 6                | 0,62 | 0,00    | 1,00 | 118-142          |
| 8)  | Satt180           | C1                            | 4                | 0,61 | 0,03    | 0,96 | 261-282          |
| 9)  | Satt294           | C1                            | 8                | 0,71 | 0,02    | 0,97 | 215-311          |
| 10) | Satt277           | C2                            | 2                | 0,38 | 0,00    | 1,00 | 190-193          |
| 11) | Satt281           | C2                            | 7                | 0,69 | 0,03    | 0,95 | 184-259          |
| 12) | Satt184           | D1a                           | 4                | 0,68 | 0,04    | 0,95 | 156-201          |
| 13) | Satt129           | D1a                           | 14               | 0,69 | 0,55    | 0,21 | 136-210          |
| 14) | Satt216           | D1b                           | 6                | 0,75 | 0,09    | 0,89 | 158-239          |
| 15) | Satt600           | D1b                           | 3                | 0,62 | 0,02    | 0,97 | 172-230          |
| 16) | Satt386           | D2                            | 5                | 0,61 | 0,06    | 0,91 | 181-224          |
| 17) | Satt389           | D2                            | 7                | 0,69 | 0,00    | 1,00 | 226-259          |
| 18) | Satt263           | E                             | 4                | 0,50 | 0,05    | 0,90 | 226-250          |
| 19) | Satt146           | F                             | 6                | 0,74 | 0,02    | 0,97 | 229-335          |
| 20) | Satt510           | F                             | 6                | 0,69 | 0,01    | 0,99 | 94-154           |
| 21) | Satt012           | G                             | 7                | 0,61 | 0,07    | 0,89 | 129-175          |
| 22) | Satt038           | G                             | 6                | 0,65 | 0,02    | 0,96 | 173-209          |
| 23) | Satt192           | H                             | 5                | 0,71 | 0,02    | 0,98 | 217-271          |
| 24) | Satt434           | H                             | 8                | 0,56 | 0,01    | 0,99 | 326-380          |
| 25) | Satt354           | I                             | 7                | 0,56 | 0,01    | 0,98 | 196-268          |
| 26) | Satt414           | J                             | 6                | 0,76 | 0,02    | 0,98 | 283-331          |
| 27) | Sat_224           | J                             | 7                | 0,74 | 0,01    | 0,99 | 140-221          |
| 28) | Satt196           | K                             | 5                | 0,26 | 0,09    | 0,65 | 190-220          |
| 29) | Satt588           | K                             | 5                | 0,27 | 0,00    | 1,00 | 137-195          |
| 30) | Satt229           | L                             | 6                | 0,44 | 0,01    | 0,98 | 209-245          |
| 31) | Satt462           | L                             | 8                | 0,77 | 0,04    | 0,94 | 250-277          |
| 32) | Satt308           | M                             | 8                | 0,81 | 0,04    | 0,95 | 147-189          |
| 33) | Satt463           | M                             | 7                | 0,67 | 0,01    | 0,99 | 116-236          |
| 34) | Satt009           | N                             | 10               | 0,73 | 0,06    | 0,91 | 180-251          |
| 35) | Satt022           | N                             | 5                | 0,70 | 0,03    | 0,96 | 215-236          |
| 36) | Satt243           | O                             | 6                | 0,39 | 0,03    | 0,93 | 219-249          |
| 37) | Satt345           | O                             | 5                | 0,57 | 0,12    | 0,79 | 218-266          |
|     | Média             | -                             | 6,0              | 0,61 | 0,04    | 0,93 | -                |

<sup>1</sup> Fontes: Cregan et al. (1999), Hwang et al. (2009) e Sayama et al. (2011).



**Figura 3.1.** Distribuição de 222 alelos identificados na genotipagem de 192 genótipos de soja, com uso de 37 locos SSR, em relação às frequências alélicas.

A estimativa do número de alelos por loco polimórfico fornece a informação de riqueza alélica no germoplasma avaliado, permitindo inferir sobre a existência de variabilidade genética e sua magnitude (Weir, 1996). Apesar dos frequentes relatos de base genética estreita na soja, advindos, dentre outras razões, do processo de melhoramento (Hiromoto & Vello, 1986; Gizlice et al., 2004; Hyten et al., 2006; Hyamanaka et al., 2007; Guo et al., 2010; Mikel et al., 2010; Priolli et al., 2010; Wysmierschi, 2010), verifica-se que no grupo dos genótipos avaliados, composto por material melhorado, há substancial nível de variabilidade alélica, demonstrado pelo número de alelos identificados. Isto contribuiu para uma boa caracterização da diversidade genética entre os grupos e indivíduos deste estudo, bem como para a avaliação de seu potencial de uso em programas de melhoramento de soja.

A maioria dos alelos com frequência abaixo de 0,10 foi observada em acessos introduzidos (Grupo A) e, ou, em linhagens provenientes de cruzamentos entre genitores introduzidos (Grupo AA), indicando que a introdução de germoplasma contribui para a ampliação da base genética, devido à introdução de novos alelos nas populações sob seleção. Resultado similar foi observado por Mulato et al. (2010), que identificaram número significativo de alelos raros ou exclusivos em acessos introduzidos. Assim, indicaram o seu uso como forma de proporcionar aumento da variabilidade genética disponível.

O índice de diversidade genética ( $h$ ) é uma medida mais apropriada para se conhecer a existência de variação em plantas autógamas, pois, nessas espécies, a variação

genética resulta da presença contínua de diferentes alelos distribuídos nos indivíduos da população, em homozigose (Weir, 1996). Neste estudo, dentre os 37 locos avaliados, a maioria (81,1%) apresentou  $h$  superior a 0,50 (Tabela 3.1), sendo, portanto, considerados como altamente informativos. Isto indica alta probabilidade de ocorrência de polimorfismo entre os acessos, nesses locos. Destacaram-se os locos Satt216, Satt414, Satt462 e Satt308, com valores de  $h$  superiores a 0,75 e entre seis e oito alelos. Outros dois importantes locos, por sua informatividade, foram Satt009 e Satt129. Estes apresentaram estimativas de  $h$  de 0,73 e 0,69, tendo permitido a identificação de dez e quatorze locos, respectivamente (Tabela 3.1). Estes locos, em soja, estão associados a QTLs (*quantitative trait loci*) responsáveis por: tamanho e forma do folíolo (Satt216 e Satt462); produtividade de grãos (Satt308 e Satt414); genes de resistência a *Phytophthora megasperma* f. sp. *glycines* (Satt414); conteúdo de oligossacarídeos solúveis e teor de sucrose (Satt462); conteúdo de proteína (Satt308); resistência a *Sclerotinia sclerotiorum* (Satt129 e Satt009); teor de isoflavonas no grão (Satt129); clorose por deficiência de ferro (Satt009); e número de flores por planta (Satt009) (Cregan et al., 1999; Soybase, 2013).

Os menores valores de  $h$  foram identificados para os locos Satt196 e Satt588, ambos no grupo de ligação K, apresentando estimativas de 0,26 e 0,27, respectivamente. Nenhum loco, dentre os 37, foi identificado como pouco informativo, evidenciando que a seleção prévia com base em resultados de outros estudos garante maior assertividade na seleção de *primers*. No presente estudo preconizou-se o uso de locos com altos valores de PIC (Hwang et al., 2009; Sayama et al., 2011), visando potencializar a identificação de polimorfismo e melhorias na eficiência da análise e na discriminação do germoplasma sob avaliação. Considerando esse mesmo critério, Vieira et al. (2009) verificaram que o uso dessa classe de marcadores (SSR), selecionados por sua informatividade, permitiu a detecção de quantidade significativa de variabilidade, mesmo entre cultivares elite das quais se esperava obter baixa variabilidade.

O coeficiente de endogamia em cultivares de espécies autógamas é normalmente esperado como sendo igual à unidade ( $f = 1,0$ ); assumindo-se as cultivares como linhagens puras ou mistura de linhagens puras (multilinhas), em homozigose nos locos para os quais os genitores divergem (Falconer & Mackay, 1996; Ramalho et al., 2001). Normalmente, essas espécies caracterizam-se por apresentar baixíssima heterozigosidade; embora possa haver grandes diferenças em tipos homozigotos (Weir, 1996), sobretudo quando se obtêm populações a partir de seleção provenientes de genitores

divergentes. No entanto, neste estudo observou-se que a estimativa de  $f$  médio, para os 192 genótipos, com uso dos 37 marcadores SSR, foi de aproximadamente 0,93 (Tabela 3.1). Desconsiderando-se, todavia, as informações para os locos com estimativas discrepantes (Satt129, Satt196 e Satt345), obtém-se  $f$  médio de 0,96, portanto, aproximadamente igual a 1,0, conforme o esperado.

Nesse contexto, para a maioria dos locos foram constatadas estimativas positivas de heterozigosidade observada, com valor médio de aproximadamente 0,04 (Tabela 3.1). Novamente, desconsiderando as informações dos locos Satt129, Satt196 e Satt345, a heterozigosidade média observada equivale a 0,02. A manutenção da heterozigosidade ao longo das gerações de seleção ocorre pelo fato de que, nem sempre, a seleção fenotípica permite a obtenção de linhagens endogâmicas totalmente homogêneas para os caracteres morfológicos e agrônômicos de interesse. Também pode ser favorecida pela ocorrência de cruzamentos naturais, que, em soja, são normalmente relatados como sendo inferiores a 1,0% (Gazzoni, 1995; Sedyama et al., 1999). O conhecimento dessa informação torna-se relevante nos casos em que esta heterozigosidade ocorre em caracteres que são objeto de seleção, pois, com isso, pode-se dificultar o processo de produção de sementes genéticas das novas cultivares.

É importante destacar que a detecção da heterozigosidade neste estudo foi possível devido ao uso dos marcadores microssatélites, que são considerados neutros ao processo seletivo (Ferreira & Grattapaglia, 1998; Morgante et al., 2002). Contudo, efeito semelhante pode ser observado em caracteres que não são considerados no processo seletivo, ou o são, mas apenas em algumas fases deste processo. Como exemplo, cita-se a reação à peroxidase em soja, que, na maioria das vezes, é avaliada apenas nas etapas finais do processo, visando à obtenção do teste de distinguibilidade, homogeneidade e estabilidade (DHE) e do valor de cultivo e uso (VCU) do material genético, para fins de proteção e registro de cultivares no Ministério da Agricultura, Pecuária e Abastecimento.

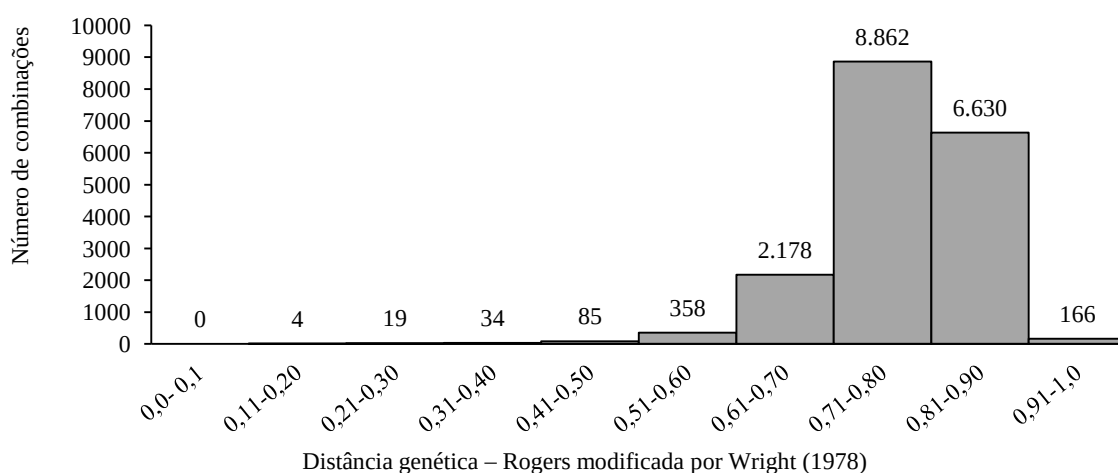
### **3.3.2 Divergência genética e estruturação populacional do germoplasma**

Os 37 locos SSR utilizados possibilitaram uma análise genética adequada do germoplasma avaliado, reforçando o grande potencial desses marcadores para uso na identificação e discriminação de genótipos (Ferreira & Grattapaglia, 1998). Com base nas estimativas de distância genética ( $d_w$ ) entre os pares de acessos, foi possível a distinção de



todas as 18.336 combinações possíveis, não tendo sido identificada qualquer combinação de acessos com distância genética igual a zero (Figura 3.2); ou seja, para os locos analisados não houve acessos idênticos (prováveis duplicatas). Apenas quatro genótipos apresentaram alta similaridade, representados pelos pares das linhagens 25 e 26 e XI 591008B(1) e XI 591008B(2). Em ambos os casos, tratam-se de linhagens isogênicas, o que confirma o alto poder de discriminação desses marcadores na identificação de genótipos redundantes. Resultados semelhantes são reportados em outros estudos, com a mesma classe de marcadores, na distinguibilidade de germoplasma melhorado, basicamente composto por cultivares comerciais (Garcia et al., 2007; Vieira et al., 2009; Priolli et al., 2010).

Do total de 18.336 combinações de distâncias genéticas, 166 pares de genótipos apresentaram estimativas superiores a 0,90 e 15.658 dessas combinações (85,39%) tiveram distâncias genéticas superiores a 0,70 (Figura 3.2). Isto representa um considerável nível de divergência genética entre os pares de genótipos, indicando bom potencial de uso do germoplasma para hibridações e obtenção de novas populações segregantes para seleção. Nessa condição, pares de acessos com elevadas estimativas de distância genética implicariam em maior probabilidade de obtenção de populações segregantes com alta variabilidade genética, potencializando as chances de sucesso no processo seletivo.



**Figura 3.2.** Distribuição do número de combinações de pares de genótipos em intervalos de distâncias genéticas, obtidas da genotipagem de 192 genótipos de soja, com uso de 37 locos microssatélites (SSR).

Com base nos dados indicados na Tabela 3.2, observa-se que a distância genética média ( $d_{wm}$ ) envolvendo todos os pares de genótipos (18.336 combinações), independentemente dos grupos de origem, foi de 0,77. Individualmente, o Grupo A apresentou a menor magnitude de  $d_{wm}$  (0,68), enquanto os Grupos B e BB, as maiores ( $d_{wm} = 0,76$ ). Considerando-se os pares de acessos de diferentes grupos de origem, os maiores valores de  $d_{wm}$  ocorreram nas combinações de genótipos envolvendo os grupos: A e B (0,80); A e BB (0,79); B e AA (0,80); e AA e BB (0,79). Com esses resultados, novamente tem-se evidência do alto potencial de geração de variabilidade desse germoplasma, sobretudo quando se consideram combinações de pares de acessos de diferentes origens (grupos) e, em especial, aquelas envolvendo introduções (A ou AA) e linhagens desenvolvidas no Brasil (B e BB).

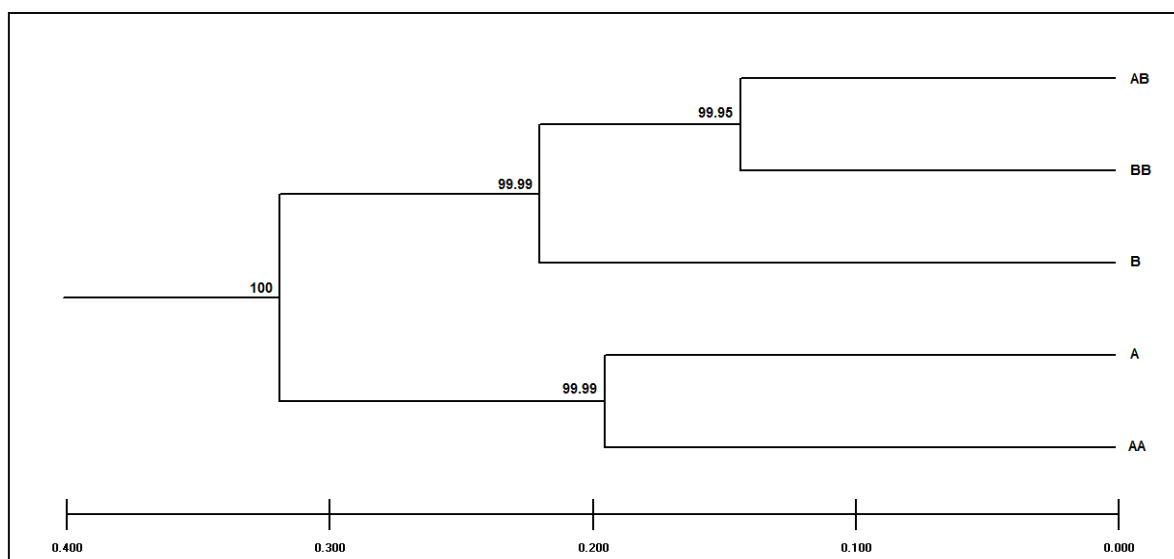
**Tabela 3.2.** Distância genética média ( $d_{wm}$ ) entre pares de genótipos (combinações) dentro e entre grupos.

| Grupos <sup>1</sup> | $d_{wm}$ | Número de combinações |
|---------------------|----------|-----------------------|
| A                   | 0,68     | 66                    |
| B                   | 0,76     | 91                    |
| AA                  | 0,73     | 595                   |
| AB                  | 0,75     | 1.953                 |
| BB                  | 0,76     | 2.278                 |
| A e B               | 0,80     | 168                   |
| A e AA              | 0,71     | 420                   |
| A e AB              | 0,77     | 756                   |
| A e BB              | 0,79     | 816                   |
| B e AA              | 0,80     | 490                   |
| B e AB              | 0,78     | 882                   |
| B e BB              | 0,77     | 952                   |
| AA e AB             | 0,78     | 2.205                 |
| AA e BB             | 0,79     | 2.380                 |
| AB e BB             | 0,77     | 4.284                 |
| Total               | 0,77     | 18.336                |

<sup>1</sup> Grupos: A- introduções; B- genótipos brasileiros; AA- linhagens obtidas de cruzamento entre genótipos introduzidos; BB- linhagens obtidas de cruzamentos entre genótipos brasileiros; AB- linhagens obtidas de cruzamentos entre genótipos introduzidos e brasileiros.

Com referência à análise de agrupamento, inicialmente consideraram-se apenas os diferentes grupos (A, B, AA, AB e BB) (Figura 3.3). Neste caso o agrupamento, no dendrograma, ocorreu predominantemente conforme a origem do germoplasma, resultando em dois grupos principais: i) grupo constituído de genótipos introduzidos (grupo A) e linhagens provenientes de cruzamentos entre genótipos introduzidos (grupo AA); e ii) grupo constituído por genótipos brasileiros (grupo B), linhagens provenientes de

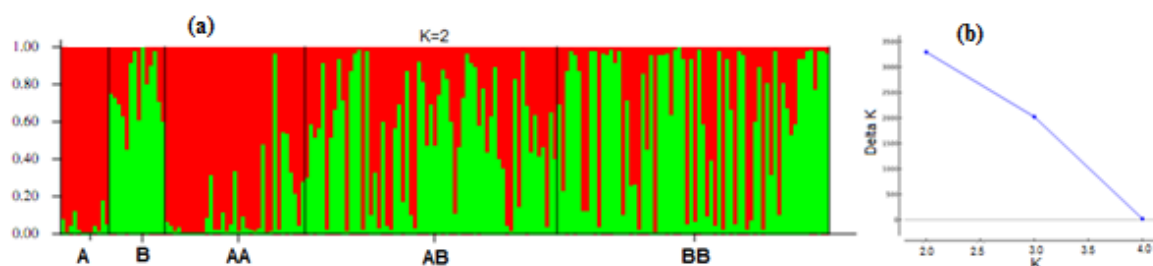
cruzamentos entre genótipos introduzidos e brasileiros (grupo AB), ou apenas entre genitores brasileiros (grupo BB). Este resultado é concordante com aquele obtido na análise de estruturação genética entre os grupos, na qual se identificou o número de duas subpopulações ( $K=2$ ) para os 192 acessos (indivíduos) avaliados (Figura 3.4). Considerando-se  $K=2$ , obteve-se um índice  $F_{ST}$  médio de 0,18, relativo à divergência entre os dois grupos. Este índice ( $F_{ST}$ ) mede a estruturação genética da população original de acessos, indicando que, da variabilidade genética total, 18% encontram-se entre essas subpopulações. Estimativas de  $F_{ST}$  de elevada magnitude, como a encontrada neste estudo, indicam que existe um processo de diferenciação entre os grupos avaliados (subpopulações), promovendo o aumento da variabilidade genética entre os grupos e a diminuição desta variabilidade dentro dos grupos (Weir, 2001). Evidencia-se, assim, a importância de se tratar os dois grupos identificados como sendo de estruturas genéticas distintas e, portanto, com frequências alélicas diferentes entre si; o que favorece o seu uso em programas de melhoramento, pois caracterizam *pools* gênicos distintos.



**Figura 3.3.** Dendrograma para cinco grupos de genótipos de soja, obtido com base na distância genética de Rogers, modificada por Wright, e 37 locos SSR, pelo método de agrupamento SAHN e critério UPGMA (correlação cofenética  $r_{\text{cof}}=0,88^{**}$ ; consistência dos nós avaliada com 10 mil *bootstraps*).

Num segundo momento, considerou-se a análise de agrupamento envolvendo todos os genótipos (acessos), independentemente de seu grupo de origem. Também neste caso observou-se um padrão de distribuição em que o agrupamento dos acessos, no dendrograma, ocorreu predominantemente conforme a estrutura de origem dos acessos, em

cada grupo. Assim, formaram-se dois grupos maiores, I e II, representados, sobretudo, por genótipos introduzidos – A (Figura 3.5) e por genótipos brasileiros – B (Figura 3.6), respectivamente.

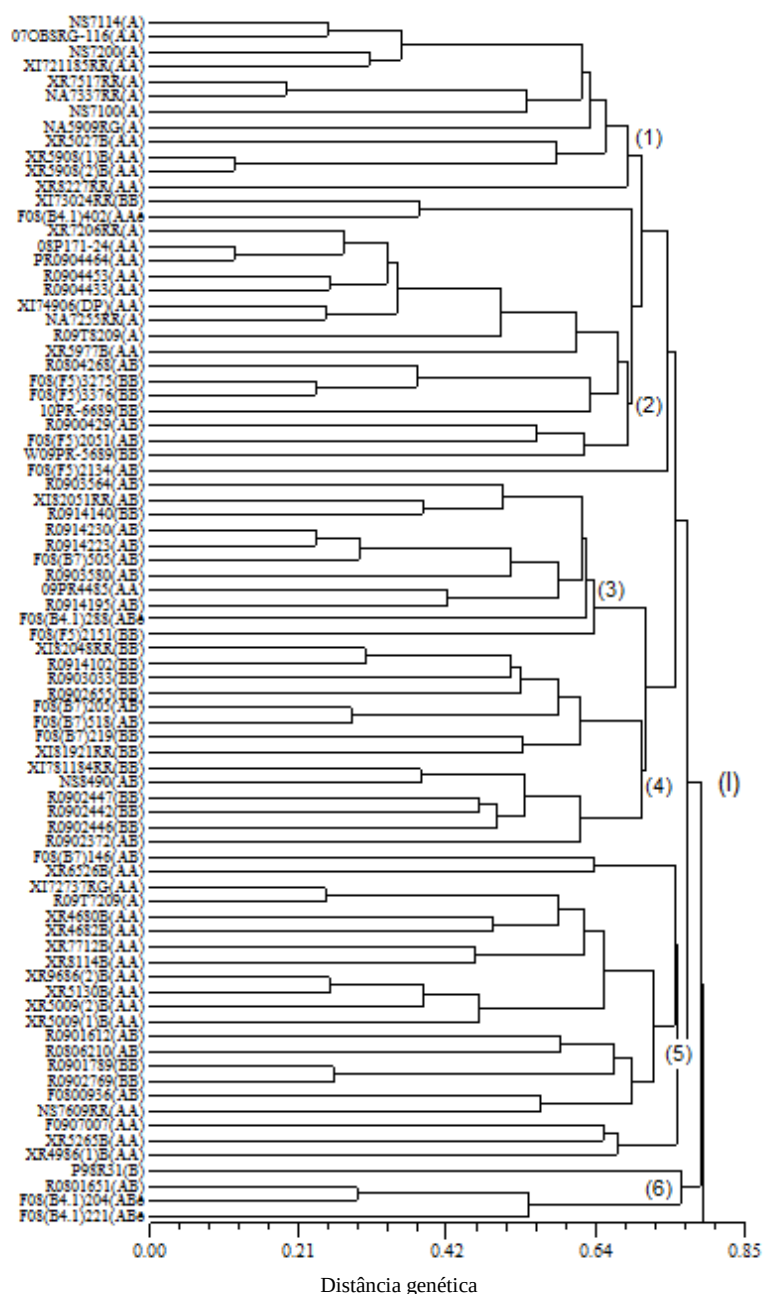


**Figura 3.4.** Estruturação genética de uma população de 192 genótipos de soja, com agrupamento prévio de cinco grupos por origem (A, B, AA, AB e BB), em duas subpopulações ( $K=2$ ; representadas por cores distintas), sendo cada acesso representado por uma linha vertical (a), e resumo da estatística  $\Delta K$  (Delta  $K$ ), sob diferentes números ( $K$ ) de subpopulações (iterações executadas com 50 mil passos na cadeia de Markov com processo de Monte Carlo – MCMC, para estabilização, seguidos de 10 milhões de passos até a obtenção das estimativas) (b).

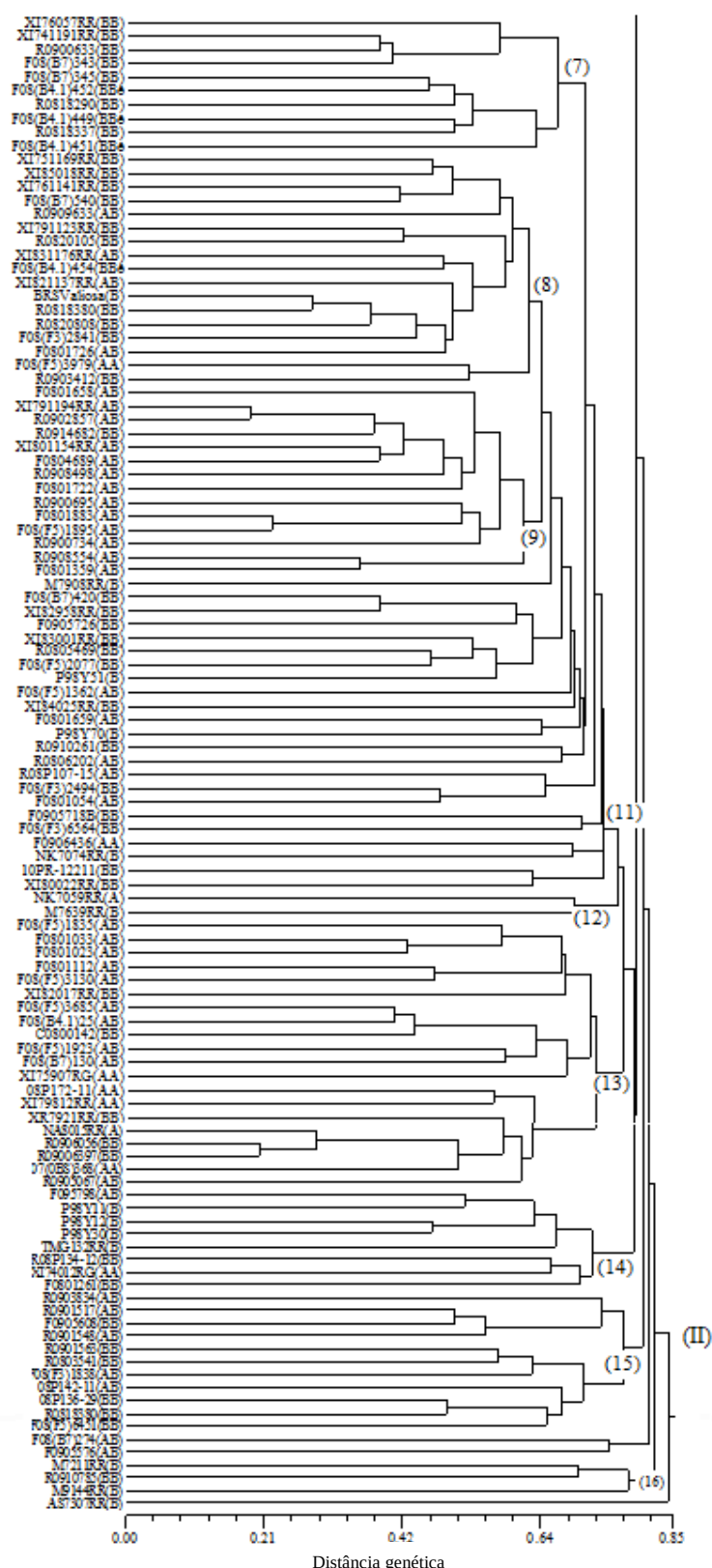
Considerando-se a primeira parte do dendrograma (Figura 3.5), observa-se que os genótipos agrupados têm origem nos grupos A, AA e AB, predominantemente. Subdividindo-se este grupo, destaca-se, no subgrupo1, a presença das cultivares NA 7337RR e NA 5909RG, e no subgrupo 2, a das cultivares NA 7255RR e Anta 8.2RR. Nos subgrupos 3 e 4, há um conjunto de linhagens, em diferentes combinações do germoplasma dos grupos A e B, destacando-se a cultivar NS 8490 no subgrupo 4, que é proveniente de cruzamento divergente entre os grupos A e B. No subgrupo 5, dois agrupamentos se caracterizam associados às cultivares BMX Potência RR e NS 7609, ambas provenientes de germoplasma introduzido. E, por fim, o subgrupo 6, que é formado pela cultivar brasileira P98R31 e outras três linhagens provenientes de combinações entre genitores dos grupos A e B.

A segunda parte do dendrograma (Figura 3.6) está relacionada, predominantemente, aos genótipos brasileiros (cultivares e linhagens endogâmicas dos grupos B e BB). Observa-se, neste caso, a presença do subgrupo predominante, identificado por 11, que possui a cultivar BRS Valiosa RR e elevado número de linhagens endogâmicas provenientes de cruzamentos em que esta cultivar foi usada como genitor. Nesse subgrupo, observam-se ainda agrupamentos menores: subgrupo 7, em que se

encontra a cultivar NS 7494; subgrupo 8, com a cultivar M 7908RR; e subgrupo 9, com as cultivares NS 8393, P98Y51 e P98Y70.



**Figura 3.5.** Dendrograma de 192 genótipos de soja (parte I), com predominância de genótipos introduzidos (grupo A), obtido com base na distância genética de Rogers, modificada por Wright, e 37 locos SSR, considerando o método de agrupamento SAHN e critério UPGMA ( $r_{\text{cof}}=0,7101^{**}$ ; consistência dos nós avaliada com 10 mil *bootstraps*).



**Figura 3.6.** Dendrograma de 192 genótipos de soja (parte II), com predominância de genótipos brasileiros (grupo B), obtido com base na distância genética de Rogers, modificada por Wright, e 37 locos SSR, considerando o método de agrupamento SAHN e critério UPGMA ( $r_{\text{cof}}=0,7101^{**}$ ; consistência dos nós avaliada com 10 mil *bootstraps*).

Ainda na Figura 3.6, observa-se o subgrupo 13, que agrupa as cultivares NS 7901 e NA 8015RR, sendo essa última proveniente de introdução de regiões temperadas, embora apresentando maior similaridade com genótipos tropicais. No subgrupo 14, destaca-se a cultivar TMG 132RR, que se agrupa às cultivares P98Y11, 98Y12 e 98Y30. No subgrupo 15 destaca-se um conjunto de linhagens provenientes de diferentes combinações entre *pools* gênicos oriundos dos grupos A e B. No subgrupo 16, encontram-se reunidas as cultivares M 7211RR e M 9144RR; e, por fim, formando um grupo de genótipo único, a cultivar AS 7307RR.

Como visto, o agrupamento com maior número de acessos corresponde ao grupo 11, que reúne linhagens e cultivares em que predomina o parentesco com a cultivar BRS Valiosa RR. Isso se deve principalmente ao fato de que, com a liberação no Brasil, em 2005, do evento GTS40-3-2, que confere tolerância ao herbicida glifosato (tecnologia RR), havia poucas fontes de germoplasma com essa característica e com adaptação a regiões tropicais brasileiras. Dentre as opções, destacava-se a cultivar BRS Valiosa RR, razão pela qual foi amplamente utilizada em programas de melhoramento, por possuir ótima adaptação a baixas latitudes, período juvenil longo, tolerância aos nematóides formadores de galhas e, principalmente, a tolerância ao herbicida glifosato. Esta cultivar foi obtida via retrocruzamentos para introgressão do gene de tolerância ao glifosato (*cp4-epsps*), a partir da cultivar BR/MG-46 (Conquista) (Arantes et al., 2005), que, historicamente, se destacou como uma das principais cultivares da soja brasileira.

O subgrupo 13, que possui as cultivares NS 7901, NA 8015RR e dezoito outras linhagens, representa genótipos de ciclo médio de maturação, com as características de tipo de crescimento indeterminado e período juvenil longo. Essa combinação de características, ainda pouco explorada e com poucas cultivares disponíveis no Brasil, tem permitido a obtenção de genótipos com ampla adaptação e elevada estabilidade produtiva nas regiões de cultivo de soja no Cerrado brasileiro. Com essa associação permite-se uma maior flexibilidade de épocas de semeadura e o cultivo em regiões de menores latitudes, sem que ocorram reduções significativas no porte da planta e, conseqüentemente, na produtividade de grãos. Além disso, as cultivares de soja com tipo de crescimento indeterminado, quando comparadas às de crescimento determinado, apresentam maior estabilidade fenotípica, principalmente em condições de estresse hídrico, como, por exemplo, sob os veranicos comuns na região central do Brasil.

O subgrupo 14 é constituído basicamente por cultivares provenientes de dois importantes programas de melhoramento desenvolvidos no Brasil, com destaque para a cultivar TMG 132RR, que atualmente é a mais plantada no Estado de Mato Grosso (mais de 1,5 milhões de hectares). As cultivares TMG 132RR, P98Y11, P98Y12 e 98Y30 respondem, aproximadamente, por 8% da área de soja plantada atualmente no Brasil (Kleffmanngroup, 2013). A característica principal desse grupo é a resistência ao nematóide de cisto (*Heterodera glycines* Ichinohe), comum a todos os acessos deste grupo. Provavelmente, essas cultivares descendem de fontes similares ou da mesma fonte de resistência à referida praga.

O subgrupo 15 também possui grande importância como germoplasma potencial, pois representa um conjunto exclusivo de linhagens de um mesmo programa de melhoramento, proveniente de fontes potenciais de resistência aos nematoides de cisto da soja (BR/MT Pintado, BRS Chapadões, entre outras) e de galhas (BRS Valiosa RR, principalmente). Além disso, mostra boa adaptação a regiões tropicais e apresenta divergência genética expressiva em relação aos demais grupos. Assim, suas linhagens são boas opções para o desenvolvimento de novas populações segregantes, objetivando a obtenção de alta variabilidade genética.

Na parte inferior do dendrograma observam-se duas linhagens formando um único grupo. Estas devem ser usadas em combinações com genótipos dos outros grupos, visando a obtenção de cruzamentos com alta variabilidade genética. Por exemplo, no subgrupo 16 estão as cultivares M 7211RR e M 9144RR, ambas pertencentes também a um mesmo programa de melhoramento. A cultivar M 9144RR é atualmente a de maior importância para regiões de baixas latitudes no Brasil. Na safra 2012/13, foram semeados em torno de dois milhões de hectares com esta cultivar, representando em torno de 8% do total da soja plantada no país (Kleffmanngroup, 2013). Estas áreas (baixas latitudes) incluem, principalmente, os Estados de Mato Grosso, Tocantins, Bahia, Maranhão e Piauí, que formam a região com maior potencial de expansão da cultura da soja no Brasil, e, de certa forma, no mundo.

Com base na análise do dendrograma completo (Figuras 3.5 e 3.6), pode-se constatar que há uma distribuição consistente dos acessos em dois grupos principais (I e II), sendo I com predominância do *pool* gênico introduzido e II com acessos desenvolvidos em programas brasileiros de melhoramento. Em ambos, observa-se a possibilidade de identificação de grupos de genótipos com diferenças genéticas entre si, o que permite o



planejamento de cruzamentos visando favorecer a obtenção de populações segregantes com alta variabilidade genética. Ademais, mesmo tratando-se de genótipos melhorados e adaptados às condições de cultivo brasileiras, verifica-se que o nível de variabilidade genética se mantém expressivo e similar ao de outros estudos em que se compararam genótipos não adaptados e introduzidos de *pools* gênicos distintos (Yamanaka et al., 2007; Mulato et al., 2010). Isso abre perspectivas para a possibilidade de introdução de genótipos elite, mantendo-se bons níveis de variabilidade genética, o que, provavelmente, facilitará a obtenção de populações segregantes com maior valor adaptativo, e, portanto, mais adequadas à seleção de genótipos promissores.

Ainda com referência ao grupo dos genótipos brasileiros, observa-se que, apesar de formarem um único grupo maior (Figura 3.3 e Figura 3.6), dentro deste há um padrão de distribuição, no qual aqueles provenientes de cada programa de pesquisa tenderam a se agrupar. Portanto, caracterizam subgrupos reais de menor distância genética entre si. Nessa condição, uma estratégia razoável para seleção de genitores seria adotar amostras provenientes de diferentes programas de melhoramento, extrapolando, com isso, a possibilidade de amostrar distintos *pools* gênicos e, obviamente, gerar maior variabilidade genética potencial. Resultado similar foi observado por Priolli et al. (2004), que avaliaram cultivares desenvolvidas por diferentes programas de melhoramento no Brasil. Contudo, esses autores concluíram que a maior variabilidade genética encontra-se intragrupos, quando comparada com aquela obtida intergrupos. Também pode-se indicar que, dentre as cultivares avaliadas no presente estudo, as cultivares AS 7307RR, M 9144RR e M 7211RR (grupo 16), NK-7059 e M 7639RR (grupo 12) caracterizam-se como as de maior potencial de uso como genitores; principalmente se forem consideradas suas combinações com genótipos do grupo I, pois, com isso, será possível uma melhor exploração potencial da variabilidade disponível.

Especificamente, os genótipos introduzidos (grupo A) e suas linhagens endogâmicas derivadas (grupo AA) representam germoplasma de *pool* gênico distinto daquele que se encontra no Brasil (Figuras 3.3 e 3.4). Estes, na maioria, são genótipos oriundos de programas argentinos de melhoramento, estabelecidos na década de 1990, e que permitiram ganhos genéticos anuais entre 1,2% e 2,0% para produtividade de grãos (Ferrarotti, 2004; Rossi, 2004). Na Argentina, a maioria destes apresenta grupo de maturidade relativa (GMR) VII e VIII. Nesses grupos de maturação, o ganho genético médio para produtividade de grãos foi de 2,0% ao ano (Ferrarotti, 2004). Nos grupos GMR

VII e VIII, desde o princípio desses programas de melhoramento, tem-se buscado desenvolver cultivares com as seguintes características: tipo de crescimento indeterminado; alta resistência ao acamamento; plantas compactas, com menor porte e maior número de nós reprodutivos; plantas com maiores índices de colheita; plantas com sementes de maior tamanho, peso e densidade; e alta produtividade de grãos (Rossi & Sartori, 2009). Aliado a isto, essas introduções têm apresentado, nas condições brasileiras, ciclos de maturação superprecoce ou precoce (96 a 115 dias), favorecendo a seleção para precocidade.

Evidencia-se que, com a introdução de germoplasma melhorado, contemplando características de interesses complementares e de bom desempenho produtivo, é possível a introdução de novos alelos, aumentando-se a base genética da espécie. Possibilita-se também a seleção de genitores com maior divergência genética, proporcionando maior probabilidade de recombinação e posterior seleção de genótipos superiores. Portanto, o uso dessa estratégia, desde que mantida a variabilidade genética, é mais vantajosa que a tradicional introdução de germoplasma não adaptado e não melhorado. Isto, ao contrário, introduziria também características indesejáveis como ausência de período juvenil longo, suscetibilidade a doenças, ocorrência de deiscência natural de vagens, maturação desuniforme, além de baixa produtividade de grãos, devido à falta de adaptação às condições de cultivo brasileiras.

Nesse contexto, o *pool* gênico representado pelas cultivares e linhagens dos grupos A e AA contribui efetivamente para a sojicultura brasileira e tem grande potencial de uso como germoplasma no país. Isto se evidencia por estimativas atuais que indicam cerca de 40% da área total cultivada no Brasil, na safra 2012/13, semeados com cultivares provenientes de genética introduzida (Kleffmanngroup, 2013). Especificamente, as cultivares NA 5909RG, NK 7059RR, BMX Potência RR, Anta 8.2RR, NA 7255RR e NA 7337RR foram plantadas em cerca de 25% dessa área. Alguns desses genótipos foram introduzidos após o lançamento no país de origem – Argentina (NA 5909RG, BMX Potência RR e Anta 8.2RR), enquanto outros foram introduzidos como linhagens em fases preliminares dos programas nacionais. Então, passaram por seleção prévia em fases intermediárias, progredindo até os testes finais nas condições de ambiente brasileiras (NA 7255RR, NA 7337RR, NK 7059RR, NS 7227, NS 7490, dentre outras). Estas se encontram entre as principais cultivares do Brasil na atualidade (Kleffmanngroup, 2013). Apresentam a particularidade de serem de ciclo precoce, com tipo de crescimento indeterminado, tendo, ainda, bom crescimento das plantas em regiões de baixas latitudes,

proporcionado pela presença de genes que conferem adequado grau de insensibilidade às variações do fotoperíodo.

A importância da soja com tipo de crescimento indeterminado tem crescido rapidamente no Brasil. Nos últimos anos, considerando-se a região Sul do país, o uso de cultivares com tal característica passou de 19%, na safra 2007/08, para em torno de 92% na safra 2011/12 (Kleffmanngroup, 2013; Kiihl, 2012). Na região central, as principais cultivares atualmente semeadas apresentam esta característica, normalmente associada aos ciclos de maturação superprecoce e precoce (ciclos entre 96 e 115 dias), além de alto potencial produtivo. O crescimento na demanda por cultivares de ciclo precoce foi impulsionado, dentre outras razões, pela dificuldade de controle da ferrugem asiática da soja, em cultivares de ciclo médio ou longo. Aliado a isto veio a crescente necessidade de exploração de cultivo em segunda safra (“safrinha”), como alternativa para se otimizar o uso das propriedades agrícolas, buscando-se maior rentabilidade do setor produtivo.

Por fim, reitera-se o potencial de uso desse germoplasma para o desenvolvimento de populações segregantes, com fins de seleção de novas cultivares. Sobretudo porque dentre as estratégias para tal desenvolvimento encontra-se o planejamento e realização de cruzamentos entre genótipos divergentes, com bom desempenho produtivo e que possuam características de interesse complementares entre si. Só assim, será possível a obtenção de populações segregantes com médias elevadas e alta variabilidade genética para os caracteres de interesse, favorecendo a maior probabilidade de sucesso na seleção de genótipos superiores. Enfim, evidencia-se que a combinação entre acessos de *pools* gênicos diferentes é favorável, pois possibilita a combinação de genótipos de bom desempenho produtivo e expressiva variabilidade genética.

### 3.4 CONCLUSÕES

- i) O uso de marcadores microsatélites mostra-se consistente na identificação de grupos genotípicos divergentes em soja, possibilitando o planejamento de combinações híbridas com maior variabilidade genética. Esta consistência demonstra-se pela elevada correspondência entre os agrupamentos obtidos e a origem comum dos acessos.
- ii) Os genótipos amostrados representam duas subpopulações com moderada estruturação genética, sendo, predominantemente, uma delas constituída por acessos introduzidos e a outra, por acessos brasileiros.

- iii) A introdução de germoplasma melhorado permite a incorporação de alelos necessários à ampliação da base genética da soja cultivada nos trópicos brasileiros, e, conseqüentemente, da respectiva variabilidade para uso no processo seletivo.

## 4 PREDIÇÃO DE VALORES GENOTÍPICOS EM ENSAIOS DE COMPETIÇÃO FINAL DE GENÓTIPOS DE SOJA

### RESUMO

A abordagem de modelos mistos tem-se demonstrado importante na predição de valores genotípicos, devido à sua flexibilidade de aplicação e às propriedades estatísticas de seus estimadores e preditores. Neste estudo, aplicou-se esta abordagem na análise de experimentos de competição final de genótipos de soja, visando avaliar suas potencialidades, relativamente à tradicional análise de modelos fixos. Os experimentos foram conduzidos na região central do Brasil, em 23 localidades, seguindo-se o delineamento de blocos completos casualizados, com três repetições. Os tratamentos foram 111 genótipos, de três grupos de maturação (superprecoce, precoce e semiprecoce). Na análise dos dados (produtividade de grãos) adotaram-se, como métodos de estimação e predição, as abordagens de modelo fixo e de modelos mistos; neste último caso, assumindo genótipos como efeitos aleatórios. Nas análises de modelos mistos, foram consideradas as seguintes modelagens para a matriz de covariâncias dos efeitos aleatórios ( $G$ ): ausência de informação de parentesco entre os genótipos; uso da informação de parentesco obtida por genealogia; e uso da informação de parentesco obtida por marcadores moleculares microsatélites (SSR). Também, para cada modelo, no contexto de modelos mistos e análise espacial, adotaram-se distintas estruturas para a matriz de covariâncias residuais ( $R$ ), buscando-se acomodar a condição de erros correlacionados. A escolha do modelo com melhor ajuste aos dados foi realizada considerando-se o teste da razão de verossimilhanças ( $LRT$ ) e os critérios de informação de Akaike ( $AIC$ ) e Bayesiano ( $BIC$ ). A herdabilidade do caráter, em nível de médias, apresentou estimativas moderadas a altas (0,52 a 0,93), revelando bom controle da variação residual, associado à elevada variabilidade genética no caráter. Observou-se que o modelo fixo, mais simples, apresentou o melhor ajuste para os diferentes grupos de experimentos. Assim, conclui-se que: i) a análise com aplicação do modelo fixo é adequada para estimar efeitos genotípicos em soja, sob condições de balanceamento e ortogonalidade, normalmente atendidas nesses ensaios de competição final; ii) a inclusão da informação de parentesco, embora não tenha proporcionado melhoria no ajuste dos respectivos modelos sob herdabilidade moderada ou baixa, aumenta a precisão das predições de valores genotípicos; iii) o uso de estruturas espaciais associadas ao modelo de análise, na presença de autocorrelação nos resíduos, melhora o seu ajuste aos dados, mesmo na condição de balanceamento e sob casualização; e iv) a escolha do modelo de análise, em condições de alta herdabilidade, não altera significativamente a classificação genotípica, logo, não impacta na seleção de genótipos superiores.

*Palavras-chave:* *Glycine max*, REML/BLUP, matriz de parentesco, modelo misto, modelo fixo.

## ABSTRACT

GENOTYPIC VALUE PREDICTION IN ADVANCED VARIETY  
TRIALS IN SOYBEAN

The mixed model approach has been shown to be important in predicting genotypic values because it is flexible use and because of the statistical properties of its estimators and predictors. In the present study, this approach was used in the analysis of data from regional yield trials of soybean genotypes for comparing its efficiency with that of the traditional fixed model. The experiments were carried out in 23 locations of Central Brazil, in randomized complete block design with three replications. The treatments were 111 genotypes from three maturity groups (very early, early and semi-early). In the grain yield data analysis the fixed and mixed model approaches were adopted as estimation/prediction methods, with genotype effects assumed as fixed and random effects, respectively. In the mixed model approach, the following models were considered for the random effects covariance matrix ( $G$ ): absence of relationship information among the genotypes; use of relationship information obtained from their genealogy; and use of relationship information obtained from microsatellite molecular markers (SSR). Also, different structures for the residual covariance matrix ( $R$ ) were adopted in the mixed models using spatial analysis, in order to consider the assumption of correlated errors. The model with best fitting to the data was chosen taking into account the likelihood ratio test ( $LRT$ ), and the Akaike ( $AIC$ ) and Bayesian ( $BIC$ ) information criteria. The estimates of heritability at means level were intermediate to high (ranging from 0.52 to 0.93), which indicated good control of the residual variation associated to high genetic variability in the trait. The simpler fixed model showed the best goodness of fit in the different maturity group trials. Therefore the conclusion was: i) the fixed model analysis is suitable to estimate genotypic effects in soybean under the conditions of balanced data and orthogonal design, which are normally assumed in the regional yield trials; ii) including relationship information associated to  $G$  matrix, although does not ensure increase in goodness of model fit under intermediate or low heritability, improve the precision in predicting genotypic values; iii) the use of spatial structures associated to  $R$  matrix, under residual autocorrelation, improved data fit, even under balanced conditions and with randomized design; and iv) the choice of the analysis model, under high heritability conditions, did not significantly change the rank of the genotypes, and, therefore, did not impact on the selection of the superior genotypes.

*Key words:* *Glycine max*, REML/BLUP, relationship matrix, mixed model, fixed model.

## 4.1 INTRODUÇÃO

O procedimento tradicional da análise de variância, derivado da teoria de quadrados mínimos, assume uma estrutura pré-definida de delineamento experimental, sendo de uso comum na análise de experimentos com plantas anuais. Ganha aplicação bastante difundida nos ensaios varietais, por exemplo, nos experimentos de competição de

linhagens, nas etapas finais dos programas de melhoramento. Nessa abordagem, a estimação de valores genotípicos é efetuada no contexto de modelo linear fixo, que tem como principal vantagem a parcimônia do método, por ser de fácil execução, demandar procedimentos computacionalmente simples, além do caráter tradicional de seu uso. Ademais, o grande apego ao método por melhoristas de plantas anuais está em que, geralmente, os experimentos de competição de genótipos são conduzidos de modo a atender às pressuposições de balanceamento dos dados, ortogonalidade do delineamento experimental e independência dos erros. Assim, justifica-se, na maioria das vezes, assumir os efeitos de tratamentos genéticos como fixos e não correlacionados.

Diversos autores relatam que, nessas condições, as estimativas obtidas pelo método de quadrados mínimos ordinários (OLS) são similares ou muito próximas às obtidas por outros métodos mais eficientes de estimação do valor genotípico, como, por exemplo, quadrados mínimos ponderados ou generalizados (Searle et al., 2002; Mrode, 2006; André, 1999; Duarte, 2000; Silva Filho, 2004; Resende, 2002; Resende, 2007). Essa equivalência entre as estimativas, considerando as suposições anteriormente citadas, é mais realista quando se estimam efeitos genotípicos em caracteres de controle genético mais simples e, ou, quando ocorre alta variabilidade genética aditiva; isto é, quando geralmente estão associadas a coeficientes de herdabilidade de elevada magnitude (André, 1999; Bauer et al., 2006; Nunes et al., 2008; Lorenzana & Bernardo, 2009).

Um agravante na estimação de valores genotípicos é introduzido quando se trabalha com caracteres quantitativos, que se caracterizam por serem controladas por vários genes de efeitos menores, com grande influência do ambiente na sua expressão fenotípica, e, conseqüentemente, baixa herdabilidade (Falconer & Mackay, 1996). Nessas condições, uma série de estudos tem demonstrado que a escolha do método de estimação assume grande importância (Bernardo, 1994; André, 1999; Duarte, 2000; Duarte & Vencovsky, 2001; Silva Filho, 2004; Bauer et al., 2006; Resende, 2002, 2007; Nunes et al., 2008; Piepho et al., 2008; Kelly et al., 2009; Lorenzana & Bernardo, 2009; Felipe & Duarte, 2010; Duarte, 2010; Pinheiro et al., 2013). Isso porque diferentes métodos podem gerar estimativas de diferentes magnitudes para os valores genotípicos, influenciando o ordenamento e a seleção de genótipos, com impactos diretos no ganho genético e na eficiência da seleção.

Em geral, ordenamentos inadequados e incorretos de genótipos, para posterior seleção, podem decorrer de inadequação das condições impostas, como, por exemplo,

assumir a irreal suposição de que a variância fenotípica é totalmente atribuída à variância genética (Resende, 2007). Sob tal suposição, assume-se a existência do coeficiente de herdabilidade igual à unidade ( $h^2 = 1$ ), o que implicaria na equivalência entre os valores genotípico e fenotípico, indicando, erroneamente, que o caráter seria completamente herdável. Por outro lado, numa condição de ausência de variância genética, que implicaria num coeficiente de herdabilidade nulo ( $h^2 = 0$ ), um modelo inadequado poderia resultar em estimativas genotípicas falsamente distintas. Conforme Felipe & Duarte (2010), adotando-se modelos menos robustos podem-se obter ordenamentos de genótipos resultantes, exclusivamente, de efeitos não genéticos (puramente ambientais), e, conseqüentemente, uma seleção de genótipos com superioridade genética irreal, sem qualquer impacto em ganhos por seleção. Essa abordagem é intrinsecamente assumida na modelagem de estimação de parâmetros considerando modelos fixos (análise de variância clássica ou OLS).

Em geral, abordagens mais complexas como a de modelos mistos envolvem suposições experimentais mais realistas; isto é, mais comumente encontradas nas condições práticas. Por exemplo, a suposição de desbalanceamento dos dados, a ocorrência de covariância genética (parentesco) entre os tratamentos e a própria natureza aleatória intrínseca aos efeitos genotípicos têm gerado previsões de valores genotípicos de maior acurácia, possibilitando maior eficiência no processo de seleção de genótipos superiores. Conceitualmente, a modelagem via modelos mistos assume, realisticamente, a suposição de que a estimativa do coeficiente de herdabilidade encontra-se entre o valor nulo e a unidade ( $0 \leq h^2 \leq 1$ ). Esta condição, sem dúvida, é genericamente mais adequada à realidade das populações sob avaliação em melhoramento de plantas (Resende, 2007; Piepho et al., 2008; Duarte, 2010).

Vale ressaltar que na abordagem de modelos mistos assumindo efeitos de tratamentos (genótipos) como aleatórios, torna-se necessário o conhecimento e uso da matriz de variância-covariância dos efeitos aleatórios ( $G$ ), para a predição dos efeitos genotípicos. A matriz  $G$  é dada por  $G = A\sigma_g^2$ , sendo  $A$ , no contexto de melhoramento genético, uma matriz de parentesco entre os genótipos. Essa matriz possui propriedades que, uma vez incluídas no modelo, garantem melhor ajustamento dos dados em relação à média geral, via efeito *shrinkage* ou “encolhimento” proporcionado por herdabilidades inferiores a 1,0 (Bernardo, 2002; Duarte, 2000; Duarte & Vencovsky, 2001; Resende 2002,



2007; Piepho et al., 2008; Felipe & Duarte, 2010). Isto corresponde a um fator de confiabilidade aplicado às médias fenotípicas, visando eliminar efeitos residuais de ambiente, numa abordagem quase sempre mais realista e de maior precisão (Resende & Duarte, 2007).

Na prática, a modelagem estatística nessa abordagem visa, principalmente, a obtenção dos BLUE (*best linear unbiased estimation*), para os efeitos fixos no modelo, e dos BLUP (*best linear unbiased prediction*), para os efeitos aleatórios. Para isso, diferentes informações associadas à matriz  $G$  podem ser introduzidas no processo de estimação/predição (Piepho et al., 2008). Pode-se considerar apenas a informação da variância genética ( $\sigma_g^2$ ) estimada entre genótipos, quando não se tem, por exemplo, informação do parentesco entre eles; o que resulta em:  $A=I$  e  $G=I\sigma_g^2$ , sendo  $I$  uma matriz identidade. Em outras situações, podem-se obter diferentes matrizes  $G$ , dadas, genericamente, por:  $G=A\sigma_g^2$ , sendo  $A$  uma matriz de coeficientes de parentesco, obtidos diretamente das genealogias dos genitores ou, indiretamente, por marcadores moleculares. Tudo isso também contribui para um cenário de pressuposições mais realistas ao contexto da análise.

O uso de modelos mistos permite, ainda, adotar modelagens com diferentes estruturas para a matriz de variância-covariância dos erros experimentais ou resíduos ( $R$ ). Neste caso, esta matriz assume uma forma geral dada por  $R=\Sigma$ , em substituição à tradicional estrutura  $R=I\sigma_e^2$ , de erros independentes, assumida na abordagem de modelos fixos. Essa é uma importante estratégia de ajuste dos dados, por análise estatística, num contexto de análise espacial, que pode favorecer a obtenção de estimativas ou predições de maior acurácia em condições de erros correlacionados (Duarte, 2000; Duarte & Vencovsky, 2005; Smith et al., 2005; Resende, 2007; Piepho & Williams, 2010; Leiser et al., 2012; Williams & Piepho, 2013). Conforme Resende (2007), a casualização e o controle local baseados em blocagem nem sempre concorrem para a neutralização dos efeitos de autocorrelação espacial nos experimentos, justificando, portanto, o uso de modelos com esse tipo de estrutura na análise de dados. Entretanto, Piepho et al. (2013) sustentam que a randomização continua sendo pré-requisito crucial para a obtenção de inferências válidas, seja num contexto convencional, seja em condições com tendência espacial e heterogeneidade de variâncias.

Com base nos aspectos comentados, são notórias as diferentes possibilidades de uso da informação gerada em experimentos conduzidos para caracterização fenotípica de genótipos. Sejam eles conduzidos em etapas iniciais ou nas fases avançadas dos programas de melhoramento, quando grande volume de experimentos, em diferentes localidades, épocas e, ou, anos, são requeridos. Assim, dentre as abordagens biométricas para estimação ou predição de efeitos genotípicos, pode-se optar por aquelas que fornecem resultados de maior confiabilidade e precisão, possibilitando a execução do processo de seleção de forma mais segura e eficiente.

Até a década de 1990, o uso da abordagem de modelos lineares mistos era praticamente restrito ao melhoramento animal e de plantas perenes, basicamente pela complexidade das estruturas experimentais e maior dificuldade em suas avaliações (White & Hodge, 1989; Mrode, 1996; Resende, 2002; Piepho et al., 2008). Isso, conseqüentemente, culminou com a necessidade de aplicação de métodos de estimação/predição de efeitos genéticos menos restritivos. Nos últimos anos, contudo, tem-se percebido que essa abordagem pode ser aplicada, com sucesso, também no melhoramento de espécies anuais; embora ainda possua pouca aceitação, de modo que sua aplicação rotineira ainda é pouco provável em curto prazo. Dentre os motivos para esse quadro, encontram-se, principalmente, a maior complexidade na execução das análises e a falsa ideia de que a metodologia não gera benefícios que justifiquem o esforço de sua aplicação, relativamente à abordagem tradicional de modelos fixos (Duarte, 2010).

Nesse sentido, considera-se justificável a iniciativa de estudos que agreguem informações comparativas entre os métodos de predição de valores genotípicos por modelos lineares mistos, considerando suas diferentes opções de aplicação; principalmente, pelos indícios de superioridade do método em relação à abordagem tradicional. Isso ganha relevância, sobretudo, em espécies já bastante melhoradas como a soja, em que os ganhos genéticos tornam-se cada vez mais difíceis. Assim, no contexto dos ensaios finais de competição de genótipos de soja, objetivou-se, neste estudo, a realização de análises biométricas com aplicação da abordagem de modelos mistos, em suas diferentes modelagens estatísticas, visando identificar aquela(s) mais adequada(s) para a estimação ou predição de valores genotípicos nessa cultura.

## 4.2 MATERIAL E MÉTODOS

### 4.2.1 Caracterização experimental

O material experimental foi representado por 111 genótipos de soja, testados na rede de ensaios de avaliação final do programa de pesquisa em soja da empresa Nidera Sementes Ltda., na safra de 2011/12. Estes ensaios são conduzidos anualmente com o objetivo de conhecer o desempenho dos genótipos nas diferentes regiões de cultivo da soja. Em média, uma linhagem com potencial para ser indicada como nova cultivar passa por três etapas (anos/safras) de avaliação final. Assim, é possível identificar aquelas com boa adaptabilidade, estabilidade e alta produtividade, para serem indicadas como novas cultivares ao sistema produtivo, e, ou, para uso como genitores (germoplasma) em hibridações visando o desenvolvimento de novas populações para seleção.

Outra característica desses ensaios é a cobertura da área de interesse do programa de pesquisa. Os ensaios abrangem localidades nos Estados de Goiás, Minas Gerais, Mato Grosso, Mato Grosso do Sul e São Paulo cobrindo toda a macrorregião sojícola 3 e a região edafoclimática 401, conforme Kaster & Farias (2012) (Figura 4.1). Toda essa região responde por cerca de 30% da área cultivada com soja no Brasil (Conab, 2014). Cada ensaio, para cada grupo de maturação (superprecoce, precoce e semiprecoce), é avaliado anualmente em até 23 locais dessa região (Tabela 4.1).

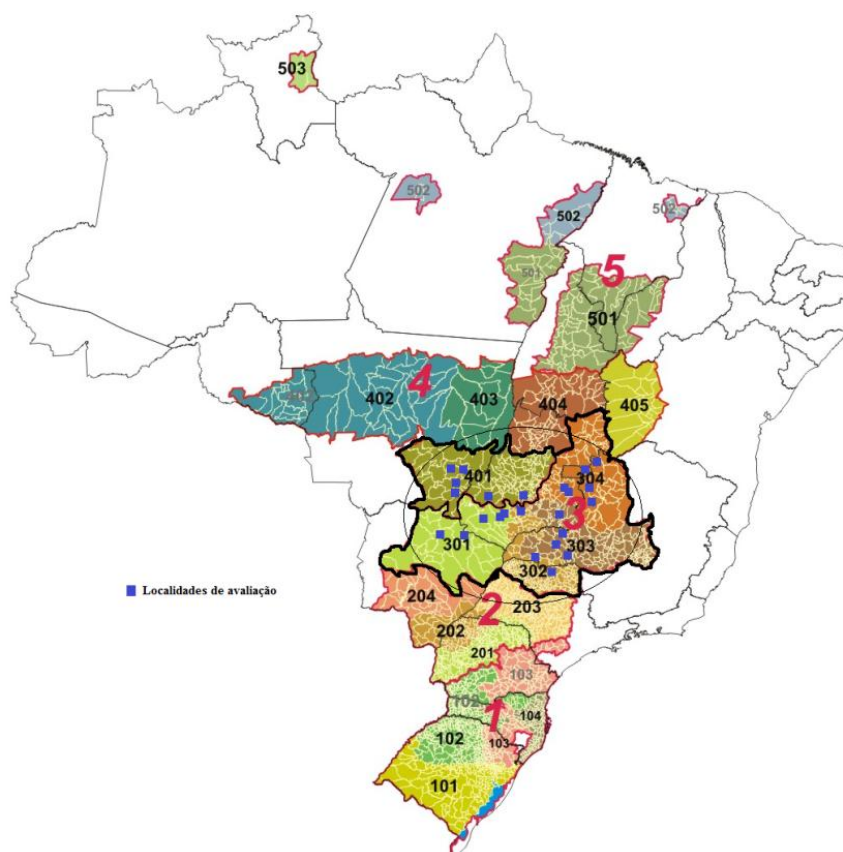
Os genótipos avaliados (tratamentos) foram constituídos por linhagens experimentais – genótipos advindos de seleção nas etapas intermediárias do programa de melhoramento (em geração  $F_8$  a  $F_{12}$ ), e por cultivares testemunhas – materiais comerciais indicados ao cultivo na região alvo do estudo. Como pré-requisito técnico para uma cultivar ser adotada como testemunha, consideram-se: boa representatividade na área de cultivo, alto potencial produtivo, boa adaptabilidade e estabilidade, e boa aceitação pelos agricultores. Legalmente, é necessário, ainda, que esteja inscrita no Registro Nacional de Cultivares do Ministério da Agricultura e Pecuária (RNC/MAPA), bem como recomendada para cultivo na região de interesse<sup>4</sup>.

Na configuração dos ensaios finais, os genótipos são classificados e alocados em conjuntos experimentais, de acordo com o seu grupo de maturação. O grupo de

---

<sup>4</sup> Informação disponível em: <[http://www.agricultura.gov.br/arq\\_editor/file/vegetal/RegistroAutorizacoes/SOJA%20\(GLYCINE%20MAX\)\(1\).doc](http://www.agricultura.gov.br/arq_editor/file/vegetal/RegistroAutorizacoes/SOJA%20(GLYCINE%20MAX)(1).doc)>.

maturação é caracterizado pelo ciclo total das plantas (da emergência à colheita), em dias, nos respectivos genótipos. Neste trabalho, avaliaram-se os seguintes conjuntos de genótipos: superprecoce (até 107 dias), precoce (108 a 115 dias) e semiprecoce (116 a 124 dias). Assim, estes conjuntos foram estruturados com 28, 40 e 43 tratamentos genéticos, respectivamente (Tabela 4.1).



**Figura 4.1.** Macrorregiões sojícolas brasileiras e distribuição geográfica das localidades de avaliação dos genótipos de soja, na Macrorregião 3 e Região edafoclimática 401 (Kaster & Farias, 2012).

Os ensaios foram conduzidos no delineamento experimental de blocos completos casualizados (DBC), com três repetições. A unidade experimental ou parcela foi constituída por quatro fileiras de plantas, com 5,0 m de comprimento, e espaçadas entre si em 0,5 m. Na coleta de dados, consideraram-se como área útil as duas fileiras centrais de cada parcela.

Nesses experimentos, os genótipos são avaliados em relação a diversos caracteres agrônômicos, incluindo produção de grãos e reação às principais doenças. Entretanto, neste estudo analisou-se apenas a produtividade de grãos, que foi obtida pela

transformação da massa de grãos (g) da área útil de cada parcela, em  $\text{kg.ha}^{-1}$ , padronizando-se para a umidade a 13%.

**Tabela 4.1.** Localidades de condução dos Ensaios de Avaliação Final (EF), na safra 2011/12, por grupo de maturação (GM), com indicação da data de semeadura, posição geográfica e número total de tratamentos genéticos e de localidades.

| Localidade            | REC <sup>1</sup> | EF/GM <sup>2</sup> |    |    | Data plantio | Posição geográfica |                 |              |
|-----------------------|------------------|--------------------|----|----|--------------|--------------------|-----------------|--------------|
|                       |                  | SUP                | P  | SP |              | Latitude Sul       | Longitude Oeste | Altitude (m) |
| Mineiros-GO           | 301              | x                  |    |    | 21/10/11     | 17°26'52,4"        | 52°48'14,8"     | 900          |
| Rio Verde-GO(1)       | 301              | x                  | x  | x  | 17/10/11     | 17°44'40,7"        | 50°57'48,8"     | 860          |
| Rio Verde-GO(2)       | 301              | x                  | x  | x  | 28/10/11     | 17°44'59,4"        | 51°00'15,8"     | 820          |
| Serranópolis-GO       | 301              | x                  | x  | x  | 18/10/11     | 18°09'10,4"        | 51°53'48,7"     | 835          |
| Chapadão do Sul-MS    | 301              | x                  | x  |    | 19/10/11     | 18°48'33,9"        | 52°35'44,6"     | 826          |
| São Gabriel Oeste-MS  | 301              | x                  | x  | x  | 20/10/11     | 19°31'01,5"        | 54°36'29,9"     | 706          |
| Acreúna-GO            | 302              | x                  |    |    | 25/10/11     | 17°16'42,5"        | 50°22'09,8"     | 640          |
| Goiatuba-GO           | 302              | x                  | x  | x  | 26/10/11     | 17°46'33,8"        | 49°59'23,5"     | 850          |
| Conceição Alagoas-MG  | 302              | x                  | x  |    | 03/11/11     | 19°46'23,3"        | 48°21'58,4"     | 612          |
| Batatais-SP           | 302              | x                  |    |    | 05/11/11     | 20°51'38,3"        | 47°28'59,5"     | 830          |
| Araxá-MG              | 303              | x                  | x  | x  | 04/11/11     | 19°27'07,6"        | 46°58'28,2"     | 1010         |
| Monte Carmelo-MG      | 303              | x                  | x  | x  | 04/11/11     | 18°55'39,5"        | 47°28'21,1"     | 1020         |
| Nova Ponte-MG         | 303              | x                  |    |    | 03/11/11     | 19°24'18,7"        | 47°42'38,1"     | 1000         |
| Cristalina-GO         | 304              | x                  | x  | x  | 29/10/11     | 16°51'19,2"        | 47°25'45,2"     | 965          |
| Luziânia-GO           | 304              | x                  |    |    | 29/10/11     | 16°29'57,7"        | 47°51'41,6"     | 871          |
| Vianópolis-GO(1)      | 304              | x                  | x  | x  | 28/10/11     | 16°51'38,9"        | 48°23'35,0"     | 972          |
| Vianópolis-GO(2)      | 304              |                    | x  | x  | 18/11/11     | 16°44'45,9"        | 48°28'14,2"     | 1005         |
| Unaí-MG               | 304              | x                  |    |    | 30/10/11     | 16°26'51,0"        | 47°12'19,2"     | 975          |
| Itaberaí-GO           | 401              | x                  |    |    | 27/10/11     | 16°03'30,5"        | 49°52'37,1"     | 760          |
| Montividiu-GO         | 401              | x                  | x  |    | 24/10/11     | 17°24'06,3"        | 51°15'34,6"     | 903          |
| Alto Garças-MT        | 401              | x                  |    |    | 26/10/11     | 16°50'35,7"        | 53°56'14,3"     | 685          |
| Campo Verde-MT        | 401              | x                  | x  | x  | 28/10/11     | 15°25'41,8"        | 55°13'23,6"     | 740          |
| Itiquira-MT           | 401              | x                  |    |    | 26/10/11     | 17°22'31,2"        | 54°22'59,2"     | 660          |
| Primavera Leste-MT    | 401              | x                  |    |    | 27/10/11     | 15°00'57,2"        | 54°12'20,7"     | 662          |
| Número de tratamentos |                  | 28                 | 40 | 43 |              |                    |                 |              |
| Número de locais      | -                | 23                 | 14 | 11 | -            | -                  | -               | -            |

<sup>1</sup> REC: região edafoclimática sojícola, conforme classificação de Kaster & Farias (2012).

<sup>2</sup> SUP: superprecoce (até 107 dias); P: precoce (108 a 115 dias); SP: semiprecoce (116 a 124 dias).

## 4.2.2 Análises biométricas

As análises biométricas da produtividade de grãos foram realizadas com ênfase na aplicação da abordagem de modelos mistos, visando elucidar suas potencialidades e vantagens. Nesse sentido, adotaram-se como método de estimação a metodologia

tradicional da análise de variância sob efeitos fixos de genótipos (modelo fixo), denominado BLUE<sub>F</sub>, e a de predição de valores genotípicos, via modelos mistos, assumindo-se genótipos como efeitos aleatórios. Nesta metodologia foram consideradas as seguintes modelagens para a matriz  $G$  (variâncias-covariâncias genotípicas): a) sem uso de informação de genealogia ou parentesco entre os genótipos – BLUP<sub>E</sub>; b) com uso da informação de parentesco entre os genótipos obtida de suas genealogias – BLUP<sub>P</sub>; e c) com uso da informação de parentesco obtida por marcadores moleculares microssatélites (SSR) – BLUP<sub>M</sub>. Também, para cada modelo mencionado, num contexto de análise espacial, adotaram-se distintas estruturas para a matriz  $R$  (variâncias-covariâncias residuais), buscando-se avaliar modelos menos restritivos sob condições de erros correlacionados, ou seja, de dependência espacial nos dados ( $R = \Sigma$ ).

#### 4.2.2.1 Estimação de efeitos genotípicos por modelo fixo

Inicialmente, aplicou-se o procedimento tradicional da análise de variância, que permite a estimação de médias genotípicas. O modelo estatístico adotado, conforme essa abordagem (quadrados mínimos ordinários) e considerando-se a análise conjunta em blocos completos casualizados, é dado por:

$$y_{ijk} = m + b_{(j)k} + g_i + a_j + (ga)_{ij} + e_{ijk} \quad (1)$$

em que:  $y_{ijk}$  é o valor fenotípico do tratamento  $i$  no bloco  $k$  dentro do local  $j$ ;  $m$  é a média geral;  $b_{(j)k}$  é o efeito de bloco, aninhado em local;  $g_i$  é o efeito de tratamento ou genótipo;  $a_j$  é o efeito de local ou ambiente;  $(ga)_{ij}$  é o efeito da interação de genótipo e ambiente;  $e_{ijk}$  é o erro experimental associado à observação  $y_{ijk}$ . Neste modelo (1) assumem-se que os efeitos são independentes e  $e_{ijk} \sim N(0, \sigma_e^2)$ .

Na representação matricial, a análise com modelos fixos é dada por:  $y = X\beta + \varepsilon$ , em que  $y$  é o vetor de observações,  $\beta$  é o vetor paramétrico de efeitos fixos desconhecidos, com matriz de incidência  $X$  conhecida;  $\varepsilon$  é um vetor aleatório de erros, com média zero e matriz de co-variâncias  $I\sigma_e^2$ , sendo  $\sigma_e^2$  a variância dos efeitos residuais e  $I$  uma matriz identidade. Logo,  $\text{var}(y) = V = \text{var}(\varepsilon) = R = I\sigma_e^2$ . Nesta condição o vetor de médias genotípicas é uma função paramétrica dos efeitos em  $\beta$ , do tipo  $L'\beta$ , em que a melhor estimativa linear não viciada (BLUE) pode ser obtida via expressão:

$BLUE(L' \beta) = L' \hat{\beta}$ , em que  $L'$  é uma matriz de coeficientes dos efeitos fixos nesta função (Felipe & Duarte, 2010), e  $\hat{\beta}$ , uma solução de quadrados mínimos do respectivo sistema de equações normais ( $X'X\hat{\beta} = X'y$ ). Para efeito de comparação, as estimativas dos efeitos genotípicos obtidas por este procedimento foram denominadas de BLUE<sub>F</sub>.

#### 4.2.2.2 Predição de valores genotípicos por modelos mistos

Na abordagem de modelos mistos, o modelo estatístico adotado é descrito por:

$$y = X\beta + Z\gamma + \varepsilon \quad (2)$$

em que:  $y$  é o vetor<sub>(nx1)</sub> de observações;  $\beta$  é o vetor<sub>(px1)</sub> paramétrico de efeitos fixos, com matriz de incidência  $X$  de ordem  $n \times p$ , conhecida;  $\gamma$  é o vetor<sub>(qx1)</sub> paramétrico de efeitos aleatórios (no presente caso, excetuando-se a média, todos os demais efeitos descritos no modelo (1), i.e., um modelo aleatório ou modelo II de Eisenhart, 1947), com matriz de incidência  $Z$  de ordem  $n \times q$ , também conhecida; e  $\varepsilon$  é um vetor<sub>(nx1)</sub> aleatório de erros. Nesta abordagem, assume-se que:  $E(y) = X\beta$ ;  $E(\gamma) = E(\varepsilon) = 0$ ;  $\gamma \sim N(0, G)$ ;  $\varepsilon \sim N(0, R)$ ; e  $Var(y) = V = ZGZ' + R$ ; sendo  $G$  a matriz de variâncias-covariâncias dos efeitos aleatórios e  $R$  a matriz de variâncias-covariâncias dos erros ou resíduos.

Para solução do sistema, com a simultânea estimação dos efeitos fixos ( $\hat{\beta}$ ) e predição dos efeitos aleatórios ( $\hat{\gamma}$ ), utilizou-se o sistema de equações de modelos mistos (EMM) (Henderson et al., 1959; Henderson, 1984), dado por:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}.$$

Assim, com base nas EMM, os melhores preditores lineares não viciados (BLUP) de  $\gamma$ , que representam a solução para estes efeitos aleatórios, são obtidos pela expressão:  $BLUP(\gamma) = \hat{\gamma} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{\beta})$ . Já os melhores estimadores lineares não viciados (BLUE) associados a  $\beta$ , resultam da solução para os efeitos fixos, obtida por quadrados mínimos generalizados (Mrode, 1996; Searle et al., 1992):

$$\hat{\beta} = [X'R^{-1}X - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}]^{-1} \cdot [X'R^{-1}y - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}y].$$

Para aplicação dessa abordagem é necessário obter as matrizes  $G$  e  $R$ , às quais estão associados os valores estimados dos respectivos componentes de variância no modelo,  $\hat{\sigma}_g^2$ ,  $\hat{\sigma}_a^2$ ,  $\hat{\sigma}_{ga}^2$ ,  $\hat{\sigma}_{b/a}^2$  e  $\hat{\sigma}_e^2$ . Para isso, obtiveram-se as estimativas dos componentes de variâncias pelo método de máxima verossimilhança restrita – REML (Patterson & Thompson, 1971), assumindo-se normalidade (Apêndice A1). O método foi aplicado em associação com os algoritmos Newton-Raphson (Dempster et al., 1977) ou AI (*average information*) (Smith et al., 2005; Butler et al., 2009), implementados, respectivamente, nos aplicativos computacionais SAS (SAS Institute, 2008) e ASReml-R (Butler, 2009; Butler et al., 2009).

Definido o sistema de equações e as respectivas matrizes de delineamento, a solução do modelo misto passa a depender, exclusivamente, das matrizes de variâncias-covariâncias genética ( $G$ ) e residual ( $R$ ). As diferentes modelagens adotadas no estudo consideraram, então, variações na estrutura destas matrizes. Foram consideradas três modelagens relacionadas à definição da matriz  $G$ :

a) Sem informação de parentesco – BLUP<sub>E</sub>

Nessa abordagem, considerou-se, para a definição da matriz  $G$ , apenas a estimativa  $\hat{\sigma}_g^2$ , intrínseca aos dados. Ou seja,  $G = I\hat{\sigma}_g^2$ , em que se supõe que os genótipos não têm relação outra entre si, senão aquela determinada pelo fato de pertencerem à mesma população (origem); assumindo, portanto, a interdependência genotípica dos tratamentos associada à variância  $\sigma_g^2$  comum. Esse é o caso mais simples de BLUP, quando o efeito genotípico é obtido por um modelo linear que o considera como variável aleatória independente, sem qualquer correlação imposta por genealogia (Piepho et al., 2008). O ganho em acurácia da predição BLUP, neste caso, está relacionado apenas à propriedade *shrinkage* do preditor (Duarte & Vencovsky, 2001; Piepho et al., 2008).

b) Com informação de parentesco obtida por genealogia – BLUP<sub>p</sub>

Nas demais formas de estruturação da matriz  $G$  foram consideradas informações de parentesco entre os genótipos. Nestes, o procedimento consiste em ponderar  $\hat{\sigma}_g^2$  pela matriz de parentesco genético  $A$ , de forma a obter:  $G = A\hat{\sigma}_g^2$ . No presente caso, aqui denominado BLUP<sub>p</sub>, a matriz  $A$  foi constituída pelas estimativas do



coeficiente de parentesco de Wright ( $f_{xy} = \frac{2r_{xy}}{\sqrt{(1+F_x)(1+F_y)}}$ ), entre os genótipos avaliados, com uso da informação de genealogias. O coeficiente  $f_{xy}$  consiste na probabilidade de que dois alelos, num mesmo loco, sejam idênticos por descendência (Kempthorne, 1957; Falconer & Mackay, 1996; Bernardo, 1994; Miranda et al., 2007; Nunes et al., 2008, Piepho et al., 2008).

A obtenção da matriz  $A$  de coeficientes  $f_{xy}$  foi calculada diretamente no programa ASReml-R, com uso do comando “asreml.Ainverse()”, considerando-se a informação das genealogias entre os genótipos (Butler, 2009; Butler et al., 2009) e a estimativa do coeficiente de endogamia médio ( $f=0,93$ ) obtida em estudo preliminar (seção 3 deste trabalho). De posse dessa matriz, que foi usada na ponderação da estimativa  $\hat{\sigma}_g^2$ , a solução de EMM resultou nas previsões BLUP<sub>P</sub> dos valores genotípicos.

#### c) Com informação de parentesco obtida por marcadores moleculares – BLUP<sub>M</sub>

O uso de marcadores moleculares na obtenção da matriz  $A$  é potencialmente superior ao uso da genealogia, pois permite obter a contribuição genômica atual de um dado genitor, para cada indivíduo relacionado (Piepho et al., 2008). Também, com o uso de marcadores, obtém-se uma estimativa mais real da porcentagem de alelos em comum entre os indivíduos (Carneiro et al., 2006); além de valores mais precisos do coeficiente de parentesco (Bernardo, 1994). Logo, essa estratégia pode viabilizar a obtenção da matriz  $A$  mesmo quando não se conhece a origem do material genético. Assim, com uso de informação de marcadores moleculares microssatélites (SSR), obteve-se a matriz de parentesco  $A$ , a qual foi considerada na ponderação de  $\hat{\sigma}_g^2$ , para compor  $G = A\hat{\sigma}_g^2$ . Seguiu-se, então, à solução de EMM para obtenção das previsões genotípicas denotadas por BLUP<sub>M</sub>.

Merece detalhamento, neste caso, o procedimento de obtenção da matriz  $A$ , a partir da informação dos marcadores codominantes SSR. Para isso, considerou-se inicialmente o método de Ritland (1996), que resulta na estimação de  $\rho_i$ , o coeficiente de parentesco (entre pares de genótipos) ou de endogamia (individual), para cada loco:  $\hat{\rho}_i = (S_i - P_i^2)/(P_i Q_i)$ , em que  $S_i$  é a proporção observada de pares similares para o alelo  $i$ ; e  $P_i = 1 - Q_i$  é a frequência estimada do alelo  $i$  para cada loco. Após, obteve-se a estimativa  $\hat{\rho}$  de parentesco ou endogamia total, considerando os 37 locos SSR, dada por:

$\hat{\rho} = \sum_i w_i \hat{\rho}_i$ , sendo  $w_i$  um fator de ponderação à unidade para as estimativas de relacionamento ( $\hat{\rho}_i$ ) associadas ao alelo  $i$ , computando todos os locos amostrados. Para cada genótipo, obteve-se ainda o coeficiente de parentesco individual (intragenotípico), dado pela relação:  $\hat{r}_{x'x} = (1 + F_x) / 2$  (Falconer & Mackay, 1996), em que  $F_x$  é o coeficiente de endogamia dado pela estimativa  $\hat{\rho}$ . As estimativas obtidas por este método refletem os coeficientes de parentesco de Malécot ( $r$ ) (Ritland, 1996). Por fim, estimaram-se os coeficientes de parentesco de Wright:  $\hat{f}_{xy} = \frac{2r_{x'y}}{\sqrt{(1 + F_x)(1 + F_y)}} = 2\hat{\rho}$ , entre pares de genótipos, e  $\hat{f}_{xx} = 2r_{x'x} = 2(1 + \hat{\rho}) / 2$ , para cada genótipo; que correspondem, respectivamente, aos elementos fora da diagonal ( $\hat{f}_{xy}$ ) e na diagonal principal ( $\hat{f}_{xx}$ ) da matriz  $A$ .

#### 4.2.2.3 Flexibilização da matriz $R$ para predição de valores genotípicos

Dentre as vantagens da aplicação dos modelos mistos encontra-se a possibilidade de adoção de estruturas de variâncias-covariâncias que considerem a existência de erros correlacionados, advindos, por exemplo, da presença de autocorrelação espacial nos dados. A autocorrelação espacial é normalmente caracterizada pela similaridade ou correlação entre observações de parcelas vizinhas, culminando na não independência entre observações e, conseqüentemente, dos erros. A abordagem de modelos mistos acomoda situações como esta, considerando, também, estruturas diferentes de  $R = I\sigma_e^2$ ; assim, permitem-se análises válidas mesmo numa condição de erros correlacionados (Duarte, 2000; Resende, 2002; 2007; Leiser et al., 2012; Williams & Piepho, 2013). Neste caso, adota-se a notação  $R = \Sigma$ , em que  $\Sigma$  representa uma estrutura geral que acomoda, ou não, a condição de independência. Torna-se necessário, então, ajustar/estimar uma função de autocovariância ou autocorrelação que descreva a dependência espacial subjacente a cada experimento. Isso foi feito com apoio integrado de técnicas de geoestatística e de análise de séries temporais. Neste trabalho, avaliaram-se, para cada modelo já descrito, as seguintes estruturas relacionadas à definição da matriz  $R$ :

a) identidade ( $I$ ), que considera independência dos erros;

- b) autorregressiva de primeira ordem (AR1), no sentido de linhas e colunas (AR1 x AR1);
- c) geométrica exponencial (SP);
- d) autorregressiva com médias móveis (ARMA); e,
- e) não estruturada (UN).

#### 4.2.3 Comparações entre abordagens estatísticas

Os critérios de escolha dos modelos que melhor se ajustaram aos dados, considerando-se as diferentes estruturas de variâncias e covariâncias associadas aos efeitos aleatórios, definidas por meio das matrizes  $G$  e  $R$ , foram: o teste da razão de verossimilhanças (*LRT*, *likelihood ratio test*) e os critérios de informação de Akaike (*AIC*) (Akaike, 1974) e Bayesiano (*BIC*) (Schwarz, 1978).

O teste *LRT* compara estimativas obtidas por máxima verossimilhança, para diferentes modelos tomados dois a dois, sendo um deles uma versão restrita do outro. De acordo com Resende (2007), esse teste permite comparar modelos que tenham uma estrutura hierárquica, em relação aos fatores aleatórios e efeitos fixos. Nessa condição, um modelo tem  $r$  parâmetros adicionais ao outro, e o teste verifica se tais parâmetros melhoram significativamente o ajuste do modelo (Camarinha Filho, 2002). A estatística *LRT* é obtida por:  $LRT = 2 \log(\ell_{R2}/\ell_{R1}) = 2[\log(\ell_{R2}) - \log(\ell_{R1})]$ , em que  $\ell_{R1}$  representa a função de máxima verossimilhança associada ao modelo aninhado (ou restrito); e,  $\ell_{R2}$ , a função de máxima verossimilhança do modelo de referência. Sob normalidade (Apêndice A1), esta estatística apresenta, assintoticamente, distribuição de  $\chi^2$ , em que o número de graus de liberdade é a diferença entre o número de parâmetros do modelo aninhado e aquele do modelo de referência. Assim, verifica-se a hipótese de igualdade de ajuste dos modelos por meio da respectiva probabilidade ( $p$ -valor) associada ao teste (significância) (Butler et al., 2009).

O critério *AIC* foi obtido por:  $AIC = -2\ell_{Ri} + 2t_i$ . E, o critério *BIC* por:  $BIC = -2\ell_{Ri} + t_i \log(n - p)$ . Em ambas as expressões,  $t_i$  representa o número total de parâmetros de efeitos fixos e aleatórios no modelo;  $n$ , o número de observações; e  $p$ , o posto da matriz  $X$ . Estas estatísticas são obtidas a partir dos valores que maximizam a função de verossimilhança, dependendo também dos números de observações e de

parâmetros do modelo. Logo, é escolhido como adequado o modelo que apresentar o menor valor para tais critérios (Butler et al., 2009).

A eficiência relativa (*ER*) na estimação/predição dos valores genotípicos pelos diferentes modelos avaliados (modelo *i*), em relação ao modelo fixo ( $BLUE_F + I$ ), foi obtida por:  $ER = \frac{SED_{BLUE_F + I}}{SED_{modelo\ i}} \times 100$ , sendo  $SED_{BLUE_F + I}$  e  $SED_{modelo\ i}$  os erros padrão médios de diferenças entre estimativas e predições genotípicas nos respectivos modelos sob comparação (Leiser et al., 2012).

Ainda com o enfoque na seleção de modelos, estimaram-se a acurácia seletiva ( $r_{\hat{g}g}$ ) e a variância do erro de predição (*PEV*), envolvidas na estimação/predição de valores genotípicos, em cada modelo. Conforme Resende (2007), no contexto de modelos mistos com efeitos de tratamentos aleatórios, o estimador da acurácia seletiva, associado à estatística *PEV*, é dado por:

$$\hat{r}_{\hat{g}g} = \left( 1 - \frac{P\hat{E}V}{\hat{\sigma}_g^2} \right)^{1/2}; \text{ com}$$

$$P\hat{E}V = \left( \frac{\hat{\sigma}_e^2 / ab}{1 + (\hat{\sigma}_e^2 / ab) / \hat{\sigma}_g^2} \right).$$

Nestas expressões estão envolvidas as estimativas:  $\hat{\sigma}_g^2$  (variância genotípica);  $\hat{\sigma}_e^2$  (variância residual); além de *a*, o número de ambientes ou experimentos, e *b*, o número de repetições ou blocos<sup>5</sup>, dada a estrutura dos experimentos em blocos completos casualizados.

Com base nessas estatísticas, adota-se como modelo de melhor ajuste aquele que apresentar maior acurácia e menor *PEV*. Conforme Resende & Duarte (2007), valores de acurácia próximos à unidade são considerados como adequados. Também indicaram que, quanto menor o valor de *PEV*, maior é a acurácia e a precisão no processo de predição. Portanto, minimizando-se *PEV* conduz-se também à maximização da acurácia.

Estimou-se também o coeficiente de herdabilidade (determinação genotípica) associado ao caráter, no sentido amplo. Seus estimadores, em nível individual ou de parcelas ( $h_i^2$ ) e em nível de médias ( $h_m^2$ ), são dados por (Falconer & Mackay, 1996; Piepho & Möhring, 2007):

<sup>5</sup> Número de repetições sujeito a desbalanceamento por perda de parcelas; neste caso, “b” foi estimado pela média harmônica do número de repetições.

$$\hat{h}_i^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_{ga}^2 + \hat{\sigma}_e^2}; \text{ e}$$

$$\hat{h}_m^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_{ga}^2 / a + \hat{\sigma}_e^2 / ab}.$$

Nestas expressões, além dos termos já descritos, acrescenta-se a estimativa de  $\sigma_{ga}^2$ , a variância da interação de genótipos com ambientes.

As comparações entre as diferentes estratégias de estimação/predição de valores genotípicos foram efetuadas também por meio do coeficiente de correlação de Spearman ( $r_s$ ), considerando-se os ordenamentos dos valores genotípicos estimados ou preditos pelas diferentes abordagens estatísticas, tomadas duas a duas.

#### 4.2.4 Simulação de redução de repetições e desbalanceamento nos dados

Com o intuito de avaliar a influência do número de repetições e do nível de desbalanceamento nos dados sobre as diferentes abordagens estatísticas, incluindo as distintas estruturas associadas às matrizes de variância-covariância dos efeitos aleatórios ( $G$ ) e residual ( $R$ ), realizaram-se exercícios de simulação com o próprio conjunto de dados observados (reais). Nos diferentes conjuntos experimentais foram simulados três casos: i) blocos com duas repetições (DBC/2rep), eliminando-se uma das três repetições; ii) desbalanceamento artificial, provocado ao acaso, simulando perdas de 30% das parcelas (DBC/unb1); e iii) desbalanceamento artificial simulando 50% de parcelas perdidas (DBC/unb2), também ao acaso. Para comparação dos casos, considerou-se como referência a análise em blocos completos casualizados com as três repetições originais (DBC/3rep). O ajuste dos respectivos modelos foi avaliado considerando-se apenas o critério de informação de Akaike ( $AIC$ ), conforme descrito na seção 4.2.3.

#### 4.2.5 Aplicativos computacionais empregados nas análises

Os procedimentos computacionais aplicados nas análises biométricas desta pesquisa foram executados com uso dos aplicativos SAS – *Statistical Analysis System* (SAS Institute, 2008) e ASReml-R (Butler, 2009; Butler et al., 2009). Com a biblioteca (*packages*) “asreml”, que é uma interface do programa estatístico ASReml para o *software*

R (Butler, 2009; R Core Team, 2013), é possível ajustar modelos lineares mistos para grandes conjuntos de dados, utilizando-se o método de máxima verossimilhança restrita (REML). Isto permite o ajuste das mais diversas e complexas estruturas para as matrizes de variâncias e covariâncias  $G$  e  $R$ , e, ainda, a inclusão de matrizes de parentesco (Butler et al., 2009).

Especificamente, na obtenção dos coeficientes de parentesco e de endogamia por marcadores SSR, utilizou-se o programa Spagedi (Hardy & Vekemans, 2002). A obtenção dos coeficientes de parentesco via informação de genealogias foi implementada no programa ASReml-R, com uso do comando “asreml.Ainverse()” (Butler, 2009; Butler et al., 2009).

### 4.3 RESULTADOS E DISCUSSÃO

#### 4.3.1 Avaliação do ajuste e seleção dos modelos de análise

Na Tabela 4.2 são apresentados os resultados das estatísticas utilizadas na comparação das diferentes abordagens de análise. Estas abordagens consideraram diferentes modelos estatísticos, com distintas estruturas associadas às matrizes de variâncias-covariâncias genéticas ( $G$ ) e residuais ( $R$ ). O modelo mais simples e tomado como referência, denotado BLUE<sub>F</sub>, foi o modelo fixo tradicional, resultante de análise de variância por quadrados mínimos ordinários. Os modelos BLUP<sub>E</sub>, BLUP<sub>P</sub> e BLUP<sub>M</sub> caracterizam-se por ter suas fontes de variação supostamente como de efeito aleatório, divergindo entre si basicamente pela adoção e uso de diferentes informações associadas à matriz  $G$ , no caso: matriz identidade, matriz de parentesco obtido por genealogias e matriz de parentesco obtida por marcadores SSR, respectivamente. Quanto à matriz  $R$  adotaram-se distintas modelagens visando identificar o efeito do uso de estruturas espaciais, comparativamente ao modelo base sob  $R = I\sigma_e^2$ . Observou-se que o modelo BLUE<sub>F</sub>, mais simples, apresentou o melhor ajuste para os diferentes grupos de experimentos (superprecoce, precoce e semiprecoce). Isto é caracterizado pelas menores magnitudes dos valores associados aos critérios de informação  $AIC$  e  $BIC$ , além de maior valor da estimativa logREML, independentemente da estrutura  $R$  associada ao modelo.

**Tabela 4.2.** Critérios de seleção de modelos (*AIC* e *BIC*) e teste da razão de verossimilhanças (*LRT*) em relação ao modelo fixo geral, considerando modelagens com diferentes estruturas associadas às matrizes *G* e *R*, para dados de produtividade de grãos (kg.ha<sup>-1</sup>), em grupos de experimentos de competição de linhagens de soja com distintos grupos de maturação.

| Modelo            | Estrutura             |                       | Grupo Superprecoce |            |            |         |                         | Grupo Precoce |            |            |         |                      | Grupo Semiprecoce |            |            |         |                     |
|-------------------|-----------------------|-----------------------|--------------------|------------|------------|---------|-------------------------|---------------|------------|------------|---------|----------------------|-------------------|------------|------------|---------|---------------------|
|                   | <i>G</i> <sup>1</sup> | <i>R</i> <sup>2</sup> | Np <sup>3</sup>    | <i>AIC</i> | <i>BIC</i> | logREML | <i>LRT</i> <sup>4</sup> | Np            | <i>AIC</i> | <i>BIC</i> | logREML | <i>LRT</i>           | Np                | <i>AIC</i> | <i>BIC</i> | logREML | <i>LRT</i>          |
| BLUE <sub>F</sub> | -                     | I                     | 1                  | 16928      | 16933      | -8463   | -                       | 1             | 14324      | 14329      | -7161   | -                    | 1                 | 12518      | 12523      | -6258   | -                   |
|                   | -                     | AR1                   | 69                 | 16809      | 17162      | -8335   | 255,18**                | 42            | 14155      | 14363      | -7036   | 250,99**             | 33                | 12347      | 12505      | -6140   | 235,43**            |
|                   | -                     | ARMA                  | 115                | 17067      | 17656      | -8418   | 89,38 <sup>ns</sup>     | 70            | 14388      | 14734      | -7124   | 74,09 <sup>ns</sup>  | 55                | 12545      | 12810      | -6218   | 80,68 <sup>ns</sup> |
|                   | -                     | SP                    | 46                 | 16792      | 17028      | -8350   | 225,69**                | 28            | 14161      | 14299      | -7052   | 217,60**             | 22                | 12341      | 12446      | -6148   | 219,41**            |
|                   | -                     | UN                    | 23                 | 16821      | 16939      | -8387   | 151,22**                | 14            | 14165      | 14235      | -7069   | 185,03**             | 11                | 12362      | 12415      | -6170   | 175,93**            |
| BLUP <sub>E</sub> | I                     | I                     | 5                  | 25756      | 25784      | -12873  | -                       | 5             | 21769      | 21796      | -10880  | -                    | 5                 | 18833      | 18859      | -9411   | -                   |
|                   | I                     | AR1                   | 73                 | 25563      | 25970      | -12709  | 328,36**                | 46            | 21585      | 21833      | -10746  | 266,50**             | 37                | 18583      | 18777      | -9254   | 313,93**            |
|                   | I                     | ARMA                  | 119                | 25989      | 26652      | -12876  | 5,36 <sup>ns</sup>      | 74            | 21947      | 22346      | -10900  | 39,74 <sup>ns</sup>  | 59                | 18941      | 19251      | -9412   | 0,64 <sup>ns</sup>  |
|                   | I                     | SP                    | 50                 | 25548      | 25826      | -12724  | 297,70**                | 32            | 21591      | 21763      | -10763  | 232,84**             | 26                | 18581      | 18718      | -9265   | 293,28**            |
|                   | I                     | UN                    | 27                 | 25620      | 25770      | -12783  | 179,74**                | 18            | 21595      | 21692      | -10780  | 200,28**             | 15                | 18628      | 18707      | -9299   | 224,72**            |
| BLUP <sub>p</sub> | A <sub>G</sub>        | I                     | 5                  | 25732      | 25760      | -12861  | -                       | 5             | 21739      | 21766      | -10864  | -                    | 5                 | 18843      | 18869      | -9416   | -                   |
|                   | A <sub>G</sub>        | AR1                   | 73                 | 25540      | 25946      | -12697  | 328,32**                | 46            | 21544      | 21792      | -10726  | 276,42**             | 37                | 18589      | 18783      | -9257   | 318,13**            |
|                   | A <sub>G</sub>        | ARMA                  | 119                | 25932      | 26594      | -12847  | 27,92 <sup>ns</sup>     | 74            | 21914      | 22313      | -10883  | 37,44 <sup>ns</sup>  | 59                | 18987      | 19296      | -9434   | 35,66 <sup>ns</sup> |
|                   | A <sub>G</sub>        | SP                    | 50                 | 25522      | 25801      | -12711  | 299,76**                | 32            | 21549      | 21722      | -10743  | 243,18**             | 26                | 18588      | 18724      | -9268   | 296,97**            |
|                   | A <sub>G</sub>        | UN                    | 27                 | 25589      | 25739      | -12767  | 187,44**                | 18            | 21558      | 21655      | -10761  | 207,02**             | 15                | 18636      | 18714      | -9303   | 227,32**            |
| BLUP <sub>M</sub> | A <sub>M</sub>        | I                     | 5                  | 25728      | 25756      | -12859  | -                       | 5             | 21739      | 21766      | -10864  | -                    | 5                 | 18843      | 18869      | -9416   | -                   |
|                   | A <sub>M</sub>        | AR1                   | 73                 | 25532      | 25939      | -12693  | 331,42**                | 46            | 21551      | 21799      | -10730  | 269,74**             | 37                | 18586      | 18780      | -9256   | 320,27**            |
|                   | A <sub>M</sub>        | ARMA                  | 119                | 25945      | 26607      | -12853  | 11,16 <sup>ns</sup>     | 74            | 21941      | 22341      | -10897  | 64,70 <sup>ns</sup>  | 59                | 18998      | 19308      | -9440   | 47,63 <sup>ns</sup> |
|                   | A <sub>M</sub>        | SP                    | 50                 | 25520      | 25798      | -12710  | 298,22**                | 32            | 21554      | 21727      | -10745  | 238,28 <sup>ns</sup> | 26                | 18587      | 18723      | -9268   | 297,66**            |
|                   | A <sub>M</sub>        | UN                    | 27                 | 25587      | 25737      | -12766  | 185,30**                | 18            | 21563      | 21660      | -10764  | 201,56**             | 15                | 18634      | 18713      | -9302   | 228,37**            |

<sup>1</sup> Estruturas associadas à matriz *G*: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>). <sup>2</sup> Estruturas associadas à matriz *R*: identidade (I), autorregressiva de primeira ordem, no sentido de linhas e colunas (AR1), autorregressiva com médias móveis (ARMA), exponencial geométrica (SP) e não estruturada (UN). <sup>3</sup> Np: número de parâmetros ( $\sigma^2$ ) associados ao modelo. <sup>4</sup> ns: valores não significativos em nível de 5% de probabilidade; \* e \*\*: valores significativos em níveis de 5% ou 1% de probabilidade, respectivamente (*LRT* com distribuição de qui-quadrado).

Constatado o modelo BLUE<sub>F</sub> (fixo) como o de melhor ajuste, observou-se que o uso de estrutura espacial associada à matriz *R*, que acomoda a possibilidade de erros não independentes, proporcionou melhoria nos critérios de avaliação do modelo, em comparação ao modelo base BLUE<sub>F</sub> + I, que assume independência dos erros (Tabela 4.2). Pelo *LRT*, observou-se que o uso de estruturas espaciais do tipo autorregressiva de primeira ordem (AR1), exponencial geométrica (SP) ou não estruturada (UN) proporcionaram melhorias significativas ( $p < 0,01$ ) na capacidade de ajuste dos modelos, relativamente àquele que considera a independência dos erros. Apenas a estrutura ARMA não agregou qualquer melhoria ( $p > 0,05$ ) nas análises. Ademais, esse padrão de ajuste foi comum aos três grupos de experimentos avaliados.

Dentre os modelos que acomodam estrutura espacial, os que apresentaram melhores ajustes foram os do tipo AR1 e SP, associados ao modelo fixo (BLUE<sub>F</sub>). Entre estes, para os grupos de genótipos superprecoces e médios não houve diferenças significativas ( $p > 0,05$ , resultados não apresentados); sendo, portanto, considerados modelos de resposta similar. Já para o grupo de genótipos precoce, o teste *LRT* revelou-se significativo ( $p < 0,01$ ), em favor da estrutura AR1, a qual apresentou também menor valor de *AIC* e maior valor para a estatística logREML.

Resultado semelhante foi identificado por Williams & Piepho (2013), em séries de experimentos de competição de variedades com diferentes culturas. Os autores simularam análises com distintas estruturas espaciais, no modelo de blocos completos casualizados, concluindo que houve pequena diferença entre as estruturas adotadas. Além disso, identificaram que a estrutura autorregressiva de primeira ordem (AR1), no sentido de linhas e/ou colunas, demonstrou-se adequada na análise de série de experimentos para avaliação de cultivares, em que a tendência espacial foi evidente. Também Leiser et al. (2012), analisando dados de produtividade de grãos em sorgo, sob condições de variação espacial advinda de fertilidade heterogênea do solo, concluíram que, dentre as estruturas avaliadas, o modelo espacial AR1xAR1 foi o que capitalizou os maiores ganhos em eficiência relativa, além de apresentar o melhor ajuste pelo critério de Akaike (*AIC*).

Tomando-se a medida de eficiência relativa (*ER*) dos modelos (Tabela 4.3), novamente observa-se que modelo fixo apresentou melhor desempenho na estimação de valores genotípicos sob a condição de alta herdabilidade ( $\hat{h}_m^2 = 0,93$ ). Todavia, com herdabilidades intermediárias ( $\hat{h}_m^2 = 0,52$  e  $\hat{h}_m^2 = 0,55$ ), a modelagem mista é que apresentou



as maiores eficiências relativas. Logo, por essa estatística e sob tais condições (herdabilidades moderadas), evidencia-se a eficiência da avaliação experimental com uso da análise por modelos mistos, adotando-se efeitos de tratamentos como aleatórios, conforme amplamente reportado na literatura (André, 1999; Bernardo, 2002; Resende, 2007; Nunes et al., 2008; Piepho et al., 2008; Kelly et al., 2009; Pinheiro et al., 2013). Por outro lado, o acréscimo da informação de parentesco, seja obtida de genealogias ou por marcadores SSR, não agregou melhoria na eficiência relativa dos respectivos modelos. As maiores eficiências foram observadas nos modelos que apenas recuperam informação intergenotípica via estimativa comum da variância genética (BLUP<sub>E</sub>).

**Tabela 4.3.** Eficiência relativa (%) na predição de valores genotípicos, obtidas com modelos de análise considerando diferentes informações associadas às matrizes de variâncias-covariâncias genéticas (*G*) e residuais (*R*), para dados de produtividade de grãos (kg.ha<sup>-1</sup>), em grupos de experimentos de competição de genótipos de soja, com distintos grupos de maturação e diferentes herdabilidades.

| Modelo            | Estrutura <sup>1</sup> |          | Grupo de maturação                                    |                                     |                                         |
|-------------------|------------------------|----------|-------------------------------------------------------|-------------------------------------|-----------------------------------------|
|                   | <i>G</i>               | <i>R</i> | Superprecoce<br>( $\hat{h}_m^2 = 0,93$ ) <sup>2</sup> | Precoce<br>( $\hat{h}_m^2 = 0,52$ ) | Semiprecoce<br>( $\hat{h}_m^2 = 0,55$ ) |
| BLUE <sub>F</sub> | -                      | I        | 100,00                                                | 100,00                              | 100,00                                  |
|                   | -                      | AR1      | 108,66                                                | 109,16                              | 105,09                                  |
|                   | -                      | SP       | 106,23                                                | 104,54                              | 103,49                                  |
|                   | -                      | UN       | 100,00                                                | 100,63                              | 98,50                                   |
| BLUP <sub>E</sub> | I                      | I        | 71,30                                                 | 114,05                              | 105,97                                  |
|                   | I                      | AR1      | 76,56                                                 | 117,86                              | 122,21                                  |
|                   | I                      | SP       | 76,10                                                 | 122,70                              | 123,84                                  |
|                   | I                      | UN       | 74,40                                                 | 121,56                              | 122,63                                  |
| BLUP <sub>p</sub> | A <sub>G</sub>         | I        | 70,97                                                 | 112,78                              | 102,86                                  |
|                   | A <sub>G</sub>         | AR1      | 76,37                                                 | 116,27                              | 118,28                                  |
|                   | A <sub>G</sub>         | SP       | 76,15                                                 | 121,10                              | 120,20                                  |
|                   | A <sub>G</sub>         | UN       | 74,58                                                 | 119,41                              | 121,12                                  |
| BLUP <sub>M</sub> | A <sub>M</sub>         | I        | 70,82                                                 | 112,19                              | 103,49                                  |
|                   | A <sub>M</sub>         | AR1      | 75,98                                                 | 112,74                              | 116,98                                  |
|                   | A <sub>M</sub>         | SP       | 75,62                                                 | 116,14                              | 118,80                                  |
|                   | A <sub>M</sub>         | UN       | 74,13                                                 | 115,15                              | 118,18                                  |

<sup>1</sup> Estruturas associadas à matriz *G*: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz *R*: identidade (I), autorregressiva de primeira ordem (AR1), exponencial geométrica (SP) e não estruturada (UN). <sup>2</sup> Coeficiente de determinação genotípica ou herdabilidade sentido amplo ( $\hat{h}_m^2$ ), em nível de médias, obtido do modelo de referência BLUP<sub>E</sub> + I (modelo misto, com pressuposição de erros independentes).

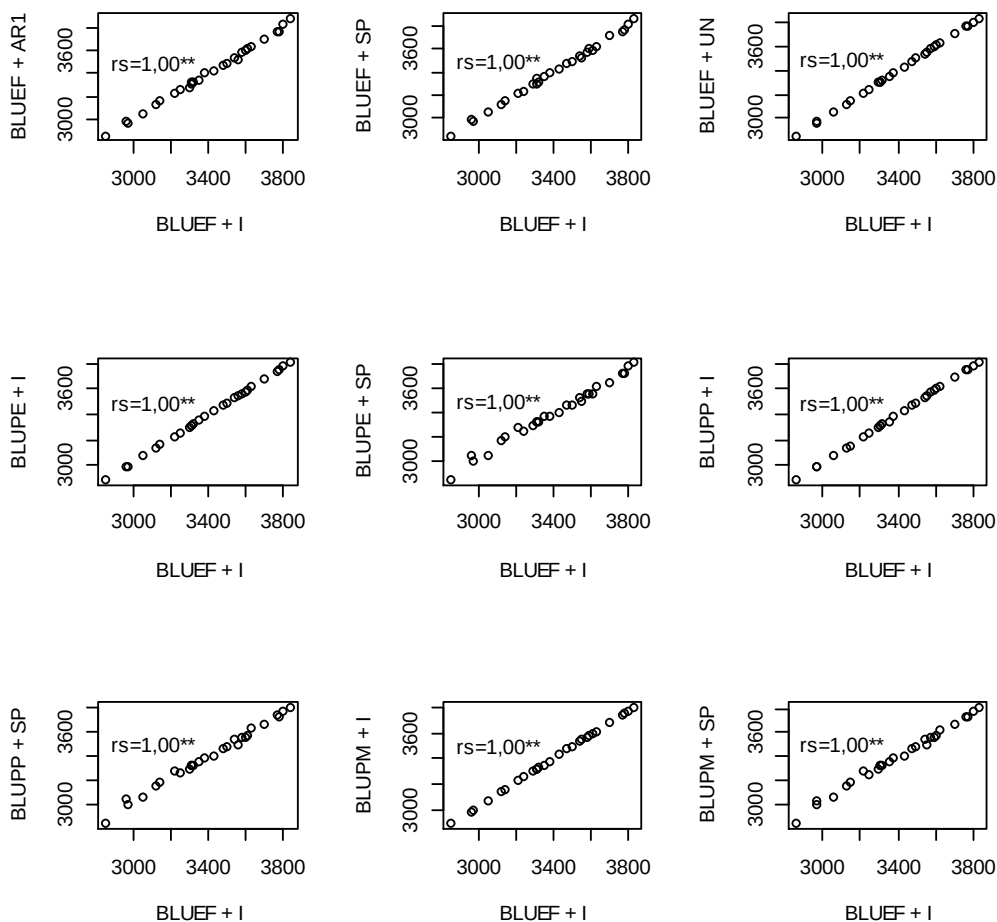
Os experimentos usados na presente pesquisa foram planejados com o intuito de se atender às pressuposições necessárias à realização da análise de modelos fixos (análise de variância clássica). Para isso, seguiu-se uma estrutura de delineamento experimental em blocos completos casualizados, com uso sistemático de repetições, que caracteriza adequado balanceamento dos dados e ortogonalidade do delineamento. Neste contexto, considerando-se os critérios de informação (*AIC* e *BIC*), os resultados indicam que o modelo fixo propiciou melhor ajuste em relação ao modelo aleatório ou misto, independentemente do uso da informação de parentesco associado à matriz *G* e, ou, da estrutura espacial associada à matriz *R*. Estes resultados reforçam a eficiência e a robustez da análise tradicional de modelo fixo na estimação de valores genotípicos, sob condições de balanceamento dos dados e uso de repetições, com controle local (blocagem).

Nessa condição, quando usadas diferentes modelagens, não se esperam alterações relevantes na obtenção das estimações/predições de valores genotípicos. Logo, não se esperam também alterações no ordenamento de genótipos por seus valores genotípicos e, conseqüentemente, em sua seleção. Assim, o uso de modelos mais parcimoniosos, ou seja, menos complexos, de certa maneira favorece a execução da análise biométrica e, portanto, devem ser preferidos. Segundo Duarte & Vencovsky (2005), modelos sobreparametrizados apenas se justificam caso demonstrem melhoria significativamente importante na qualidade de ajustamento, em relação a modelos mais simples; melhoria esta que deve ser demonstrada estatisticamente por meio de critérios comparativos como *AIC* e *BIC*, que modelos mais parametrizados. Esses critérios são obtidos a partir dos valores que maximizam a função de verossimilhança, dependendo também dos números de observações e de parâmetros do modelo. Por isso, é escolhido como mais adequado o modelo que apresente o menor valor para tais critérios (Butler et al., 2009). Tais estatísticas estão fundamentadas na teoria de decisão e penalizam os modelos com número elevado de parâmetros, buscando-se evitar excesso de parametrização e favorecendo a escolha de modelos parcimoniosos (Camarinha Filho, 2002). Piepho & Williams (2010) indicam que a minimização do critério *AIC* é uma estratégia que permite identificar um modelo aproximado, com menor discrepância de um provável modelo real, e bastante parcimonioso para ser suportado pelos dados. Apesar de ressaltarem que um modelo selecionado com base em dados observados nunca poderá ser considerado livre de erros, devendo, portanto, ser encarado como uma aproximação.

### 4.3.2 Influência da modelagem em parâmetros genéticos e na seleção genotípica

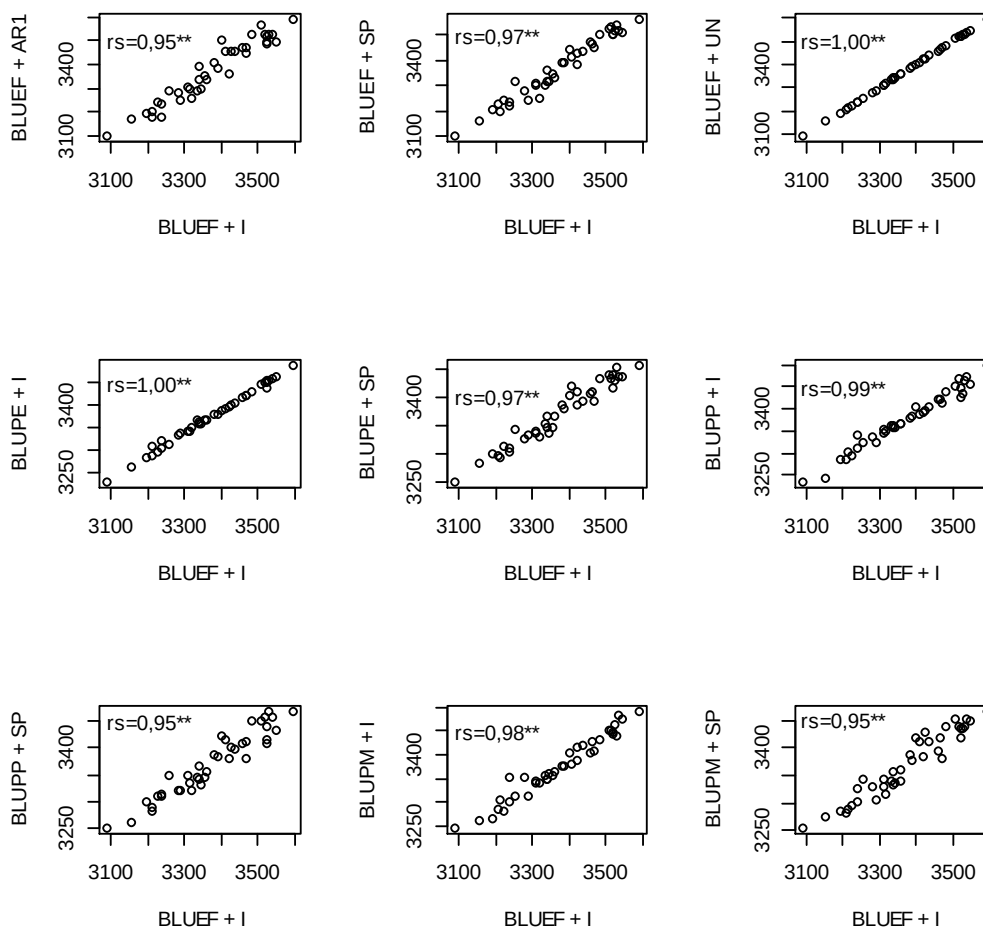
Os ordenamentos dos genótipos por seus valores genotípicos estimados ou preditos, segundo os diferentes modelos, apresentaram altos e significativos coeficientes de correlações de Spearman (correlação de postos). Evidenciou-se, assim, que nos grupos de experimentos avaliados, não houve mudanças expressivas das estimativas ou previsões obtidas por modelo fixo, em relação ao modelo aleatório ou misto com distintas estruturas de análise. Exemplificando esse efeito, observa-se que esse coeficiente de correlação, para o grupo de genótipos superprecoces, considerando o modelo fixo ( $BLUE_F + I$ ) e a modelagem mista ( $BLUP_E + I$ ,  $BLUP_P + I$  e  $BLUP_M + I$ ), foi igual à unidade ( $r_s = 1,0$ ,  $p < 0,01$ ). Note-se que estes últimos representam, respectivamente, os modelos sem uso da informação de parentesco, com uso de parentesco obtido por genealogias e com uso de parentesco obtido por marcadores moleculares SSR (Figura 4.2). Para o grupo de genótipos precoces, esse coeficiente resultou em 1,00, 0,99 e 0,98, para as mesmas respectivas comparações (Figura 4.3). No grupo de genótipos semiprecoces, muito similarmente resultou em 1,00, 0,99 e 0,97 (Figura 4.4). A ocorrência de coeficientes de correlação diferentes da unidade para os grupos de genótipos precoces e semiprecoces deve-se, provavelmente, ao desbalanceamento dos dados devido à ocorrência de parcelas perdidas nestes grupos experimentais.

O efeito observado sobre o ordenamento dos genótipos nos diferentes modelos de análise, além de favorecido pela estrutura de delineamento implementada, que permitiu o balanceamento dos dados, também foi influenciado pela elevada magnitude dos coeficientes de determinação genotípica ou herdabilidade do caráter ( $\hat{h}_m^2$ ), em nível de médias. No caso do grupo de genótipos superprecoces, que apresentou  $\hat{h}_m^2 = 0,93$  (Tabela 4.4), considerando-se o modelo misto  $BLUP_E + I$ , todas correlações com o modelo de referência ( $BLUE_F + I$ ) foram iguais à unidade ( $r_s = 1,0$ ,  $p < 0,01$ ). Isto indica o mesmo ordenamento dos genótipos nos diferentes modelos (Figura 4.2). Portanto, a escolha do modelo de análise seria indiferente nessa condição, não influenciando ou impactando na seleção dos genótipos.



**Figura 4.2.** Estimativas do coeficiente de correlação de Spearman ( $r_s$ ) para os ordenamentos de valores genotípicos preditos em produtividade de grãos ( $\text{kg} \cdot \text{ha}^{-1}$ ), de genótipos de soja do grupo de maturação superprecoce, obtidos com distintas modelagens das matrizes de variância-covariância dos efeitos aleatórios ( $G$ ) e residuais ( $R$ ), em relação às estimativas obtidas pelo modelo fixo geral.

Nos conjuntos experimentais precoce e semiprecoce, cujas estimativas  $\hat{h}_m^2$  foram 0,52 e 0,55, respectivamente (Tabelas 4.5 e 4.6), os modelos com estrutura  $R = I\sigma^2$  apresentaram alta correlação entre si, sendo esse valor diminuído apenas nos modelos com uso de estruturas que acomodam variação espacial. Contudo, mesmo nestes casos as estimativas do coeficiente de correlação foram de magnitudes elevadas e significativas, variando de 0,95 a 1,0, no grupo precoce (Figura 4.3), e de 0,88 a 1,0 no grupo semiprecoce (Figura 4.4).

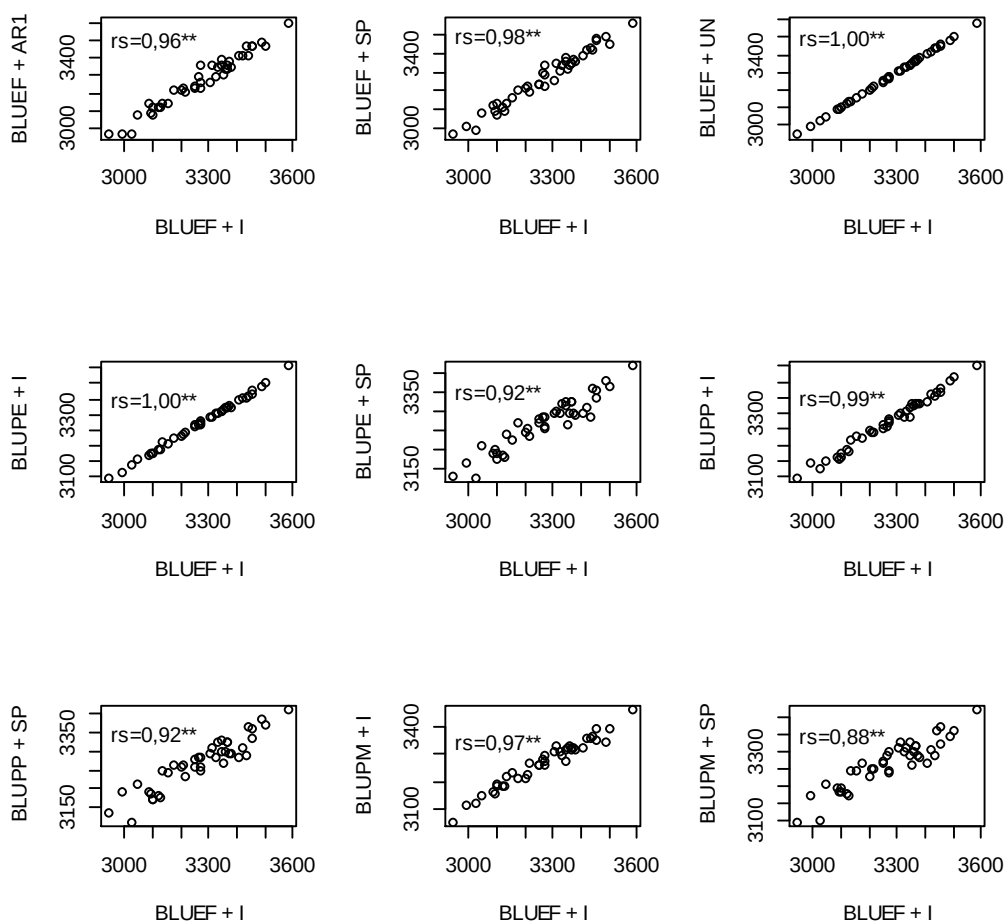


**Figura 4.3.** Estimativas do coeficiente de correlação de Spearman ( $r_s$ ) para os ordenamentos de valores genotípicos preditos em produtividade de grãos ( $\text{kg} \cdot \text{ha}^{-1}$ ), de genótipos de soja do grupo de maturação precoce, obtidos com distintas modelagens das matrizes de variância-covariância dos efeitos aleatórios ( $G$ ) e residuais ( $R$ ), em relação às estimativas obtidas pelo modelo fixo geral.

Dentre as vantagens do uso da abordagem de modelos mistos com tratamentos aleatórios está a possibilidade de se obter um estimador *shrinkage*, que representa a aplicação de um fator de confiabilidade aplicado às médias fenotípicas, visando eliminar efeitos residuais de ambiente (Resende & Duarte, 2007). Esse tipo de estimador proporciona certo “encolhimento” dos valores genotípicos em torno da média geral da população, como função direta da herdabilidade do caráter; o que pode ser notado na própria expressão geral do BLUP de  $\gamma$ :  $\hat{\gamma} = h^2(y_j - \bar{y})$  (Piepho et al., 2008).

Em outras palavras, a convergência das médias genotípicas em direção à constante dada pela média geral depende, exclusivamente, da estrutura da matriz de

variância-covariância dos efeitos aleatórios ( $G$ ), que inclui a variância genética. Logo, de forma resumida, se ocorre elevada variância genética associada à matriz  $G$ , tem-se, por conseguinte, pequeno efeito na convergência das predições. Porém, caso  $G$  assumia autovalores nulos, indicando ausência de variâncias e covariâncias genéticas entre os genótipos, obtém-se encolhimento máximo das predições genotípicas em relação à média geral. Isso significa que, no conjunto de genótipos avaliados, toda a variação é de efeito ambiental, com efeitos genotípicos nulos (Duarte & Vencovsky, 2001; Felipe & Duarte, 2010). No presente caso, como as estimativas de herdabilidade (em nível de médias) foram elevadas, esse efeito de “encolhimento” das médias  $BLUP$  em relação às respectivas médias  $BLUE$  praticamente não foi notado.



**Figura 4.4.** Estimativas do coeficiente de correlação de Spearman ( $r_s$ ) para os ordenamentos de valores genotípicos preditos em produtividade de grãos ( $\text{kg} \cdot \text{ha}^{-1}$ ), de genótipos de soja do grupo de maturação semiprecoce, obtidos com distintas modelagens das matrizes de variância-covariância dos efeitos aleatórios ( $G$ ) e residuais ( $R$ ), em relação às estimativas obtidas pelo modelo fixo geral.

Essa estratégia de análise normalmente tende a permitir a obtenção de predições com menores erros e maior acurácia, possibilitando, assim, maior assertividade e sucesso com a seleção. Contudo, conforme observado nesta pesquisa, numa condição de ortogonalidade advinda do balanceamento dos dados e do uso de repetições, numa estrutura de delineamento com controle local (blocagem) e elevadas magnitudes de herdabilidade, a influência do modelo na seleção demonstra baixa relevância. Portanto, o uso de modelos mais simples (parcimoniosos), nesta condição, mostra-se favorável; uma vez que os modelos mais complexos não permitiram ganhos significativos em acurácia e, conseqüentemente, não terão impacto sobre a seleção.

Observa-se ainda que o uso de modelos com informação de parentesco genético (BLUP<sub>P</sub> e BLUP<sub>M</sub>) ocasionou redução nas estimativas de variância genética ( $\hat{\sigma}_g^2$ ), em relação ao modelo base BLUP<sub>E</sub>, que pressupõe ausência dessa informação e independência dos erros (Tabelas 4.4 a 4.6). Conseqüentemente, isto repercutiu da mesma forma nas estimativas de herdabilidade, em nível individual ( $\hat{h}_i^2$ ) e de médias ( $\hat{h}_m^2$ ), bem como nas de acurácia seletiva ( $\hat{r}_{gg}$ ). Esse efeito foi menos pronunciado no grupo de genótipos superprecoces, que apresentou elevada magnitude para  $\hat{h}_m^2$  (0,93). Em contrapartida, foi mais pronunciado nos grupos precoces e semiprecoces, com  $\hat{h}_m^2 = 0,52$  e  $\hat{h}_m^2 = 0,55$ , respectivamente. Resultado semelhante foi observado por Pinheiro et al. (2013), ao estimarem a variância genética e a herdabilidade para produtividade de grãos em soja, comparando modelos com e sem informação de genealogias. Os autores constataram que a inclusão da informação de parentesco no modelo proporcionou reduções de 60% a 71% na variância genética aditiva, e de 13% a 25% na herdabilidade.

Segundo Resende (2007), a ocorrência de  $\hat{\sigma}_g^2$  menores em modelos mais complexos, principalmente, quando são consideradas estruturas espaciais, revela inadequação desses modelos. Isto geralmente ocorre quando as estimativas dos coeficientes de autocorrelação espacial não diferem estatisticamente da unidade. Esse fato foi observado na maioria dos modelos avaliados, nos três grupos de maturação; seja considerando as estruturas associadas à matriz  $G$ , seja considerando aquelas da matriz  $R$  (Tabelas 4.4 a 4.6). Ainda, conforme Resende (2007), esse efeito que caracteriza a existência de tendência espacial nos dados pode ser efetivamente removido pelo uso de

delineamento experimental, por meio dos efeitos de blocos e parcelas. A tendência espacial gera autocorrelações positivas entre parcelas vizinhas, advindas normalmente de diferencial de gradiente num mesmo bloco, sendo, portanto, um efeito totalmente ambiental ou residual (Duarte, 2000; Duarte & Vencovsky, 2005; Resende, 2007).

**Tabela 4.4.** Estimativas REML para os componentes de variância genética ( $\hat{\sigma}_g^2$ ), da interação de genótipos com ambientes ( $\hat{\sigma}_{ga}^2$ ) e residual ( $\hat{\sigma}_e^2$ ), coeficientes de determinação genotípica no sentido amplo, em nível individual ( $\hat{h}_i^2$ ) e de médias ( $\hat{h}_m^2$ ), e acurácia seletiva ( $\hat{r}_{gg}$ ), obtidos por modelos de análise com diferentes estruturas de informação associadas às matrizes de variância-covariância dos efeitos aleatórios (G) e residuais (R), para dados de produtividade de grãos (kg.ha<sup>-1</sup>), num grupo de experimentos de competição de genótipos de soja de maturação superprecoce<sup>1</sup>.

| Modelo            | Estrutura <sup>2</sup> |      | Componente variância |                       |                    | Coef. det. genot. |               | Acurácia<br>( $\hat{r}_{gg}$ ) |
|-------------------|------------------------|------|----------------------|-----------------------|--------------------|-------------------|---------------|--------------------------------|
|                   | G                      | R    | $\hat{\sigma}_g^2$   | $\hat{\sigma}_{ga}^2$ | $\hat{\sigma}_e^2$ | $\hat{h}_i^2$     | $\hat{h}_m^2$ |                                |
| BLUE <sub>F</sub> | -                      | I    | -                    | -                     | 155625,2           | -                 | -             | -                              |
|                   | -                      | AR1  | -                    | -                     | 160265,0           | -                 | -             | -                              |
|                   | -                      | ARMA | -                    | -                     | 137334,4           | -                 | -             | -                              |
|                   | -                      | SP   | -                    | -                     | 161107,7           | -                 | -             | -                              |
|                   | -                      | UN   | -                    | -                     | 155625,2           | -                 | -             | -                              |
| BLUP <sub>E</sub> | I                      | I    | 68160,1              | 57282,0               | 155625,2           | 0,24              | 0,93          | 0,98                           |
|                   | I                      | AR1  | 65462,4              | 53610,8               | 190989,5           | 0,21              | 0,93          | 0,98                           |
|                   | I                      | ARMA | 36309,3              | 35307,1               | 142122,9           | 0,17              | 0,91          | 0,97                           |
|                   | I                      | SP   | 64188,1              | 53402,0               | 172192,7           | 0,22              | 0,93          | 0,98                           |
|                   | I                      | UN   | 64992,5              | 53204,7               | 160241,8           | 0,23              | 0,93          | 0,98                           |
| BLUP <sub>P</sub> | A <sub>G</sub>         | I    | 43841,8              | 30305,9               | 155017,8           | 0,19              | 0,92          | 0,98                           |
|                   | A <sub>G</sub>         | AR1  | 43299,6              | 28034,7               | 170594,6           | 0,18              | 0,92          | 0,97                           |
|                   | A <sub>G</sub>         | ARMA | 33707,1              | 29492,8               | 140787,2           | 0,17              | 0,91          | 0,97                           |
|                   | A <sub>G</sub>         | SP   | 42624,2              | 27587,9               | 171972,8           | 0,18              | 0,92          | 0,97                           |
|                   | A <sub>G</sub>         | UN   | 42903,0              | 27499,5               | 159773,2           | 0,19              | 0,92          | 0,97                           |
| BLUP <sub>M</sub> | A <sub>M</sub>         | I    | 43047,8              | 27256,2               | 155025,9           | 0,19              | 0,93          | 0,97                           |
|                   | A <sub>M</sub>         | AR1  | 41361,0              | 25657,7               | 170380,7           | 0,17              | 0,92          | 0,97                           |
|                   | A <sub>M</sub>         | ARMA | 33387,3              | 28675,6               | 139449,6           | 0,17              | 0,91          | 0,97                           |
|                   | A <sub>M</sub>         | SP   | 40895,2              | 25320,6               | 171295,3           | 0,17              | 0,92          | 0,97                           |
|                   | A <sub>M</sub>         | UN   | 41144,7              | 25204,9               | 159016,4           | 0,18              | 0,92          | 0,97                           |

<sup>1</sup> Vinte e três experimentos conduzidos no delineamento de blocos completos casualizados, com 28 tratamentos e três repetições.

<sup>2</sup> Estruturas associadas à matriz G: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz R: identidade (I), autorregressiva de primeira ordem (AR1), autorregressiva com médias móveis (ARMA), exponencial geométrica (SP) e não estruturada (UN).



**Tabela 4.5.** Estimativas REML para os componentes de variância genética ( $\hat{\sigma}_g^2$ ), da interação de genótipos com ambientes ( $\hat{\sigma}_{ga}^2$ ) e residual ( $\hat{\sigma}_e^2$ ), coeficientes de determinação genotípica no sentido amplo, em nível individual ( $\hat{h}_i^2$ ) e de médias ( $\hat{h}_m^2$ ), e acurácia seletiva ( $\hat{r}_{gg}$ ), obtidos por modelos de análise com diferentes estruturas de informação associadas às matrizes de variância-covariância dos efeitos aleatórios (G) e residuais (R), para dados de produtividade de grãos (kg.ha<sup>-1</sup>), num grupo de experimentos de competição de genótipos de soja de maturação precoce<sup>1</sup>.

| Modelo            | Estrutura <sup>2</sup> |      | Componente variância |                       |                    | Coef. det. genot. |               | Acurácia<br>( $\hat{r}_{gg}$ ) |
|-------------------|------------------------|------|----------------------|-----------------------|--------------------|-------------------|---------------|--------------------------------|
|                   | G                      | R    | $\hat{\sigma}_g^2$   | $\hat{\sigma}_{ga}^2$ | $\hat{\sigma}_e^2$ | $\hat{h}_i^2$     | $\hat{h}_m^2$ |                                |
| BLUE <sub>F</sub> | -                      | I    | -                    | -                     | 190923,7           | -                 | -             | -                              |
|                   | -                      | AR1  | -                    | -                     | 174825,0           | -                 | -             | -                              |
|                   | -                      | ARMA | -                    | -                     | 131723,0           | -                 | -             | -                              |
|                   | -                      | SP   | -                    | -                     | 195421,2           | -                 | -             | -                              |
|                   | -                      | UN   | -                    | -                     | 190418,3           | -                 | -             | -                              |
| BLUP <sub>E</sub> | I                      | I    | 7656,9               | 36800,3               | 189956,3           | 0,03              | 0,52          | 0,79                           |
|                   | I                      | AR1  | 7422,7               | 42826,2               | 189230,8           | 0,03              | 0,50          | 0,78                           |
|                   | I                      | ARMA | 3867,5               | 42929,3               | 132062,2           | 0,02              | 0,38          | 0,74                           |
|                   | I                      | SP   | 6187,0               | 42631,6               | 189221,8           | 0,03              | 0,45          | 0,76                           |
|                   | I                      | UN   | 6405,4               | 40521,8               | 186829,7           | 0,03              | 0,47          | 0,76                           |
| BLUP <sub>P</sub> | A <sub>G</sub>         | I    | 4719,5               | 21379,1               | 186076,6           | 0,02              | 0,44          | 0,71                           |
|                   | A <sub>G</sub>         | AR1  | 4851,7               | 23317,6               | 187184,2           | 0,02              | 0,44          | 0,72                           |
|                   | A <sub>G</sub>         | ARMA | 2707,2               | 35427,3               | 131951,0           | 0,02              | 0,32          | 0,67                           |
|                   | A <sub>G</sub>         | SP   | 4057,3               | 23328,5               | 186980,1           | 0,02              | 0,40          | 0,68                           |
|                   | A <sub>G</sub>         | UN   | 4224,5               | 22377,5               | 183981,9           | 0,02              | 0,41          | 0,69                           |
| BLUP <sub>M</sub> | A <sub>M</sub>         | I    | 4228,8               | 18525,1               | 186752,5           | 0,02              | 0,42          | 0,69                           |
|                   | A <sub>M</sub>         | AR1  | 4478,7               | 21530,2               | 186960,9           | 0,02              | 0,43          | 0,70                           |
|                   | A <sub>M</sub>         | ARMA | 2707,2               | 33817,2               | 131707,7           | 0,02              | 0,33          | 0,67                           |
|                   | A <sub>M</sub>         | SP   | 3831,0               | 21474,3               | 186681,8           | 0,02              | 0,39          | 0,67                           |
|                   | A <sub>M</sub>         | UN   | 4018,4               | 20272,6               | 183799,6           | 0,02              | 0,41          | 0,69                           |

<sup>1</sup> Quatorze experimentos conduzidos no delineamento de blocos completos casualizados, com quarenta tratamentos e três repetições.

<sup>2</sup> Estruturas associadas à matriz G: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz R: identidade (I), autorregressiva de primeira ordem (AR1), autorregressiva com médias móveis (ARMA), exponencial geométrica (SP) e não estruturada (UN).

O caráter produtividade de grãos, estudado nesta pesquisa, é tipicamente conhecido por sua natureza quantitativa (Falconer & Mackay, 1996; Ramalho et al., 2001). Contudo, observou-se que sua seleção pode ser facilitada considerando os coeficientes de

**Tabela 4.6.** Estimativas REML para os componentes de variância genética ( $\hat{\sigma}_g^2$ ), da interação de genótipos com ambientes ( $\hat{\sigma}_{ga}^2$ ) e residual ( $\hat{\sigma}_e^2$ ), coeficientes de determinação genotípica no sentido amplo, em nível individual ( $\hat{h}_i^2$ ) e de médias ( $\hat{h}_m^2$ ), e acurácia seletiva ( $\hat{r}_{gg}$ ), obtidos por modelos de análise com diferentes estruturas de informação associadas às matrizes de variância-covariância dos efeitos aleatórios (G) e residuais (R), para dados de produtividade de grãos (kg.ha<sup>-1</sup>), num grupo de experimentos de competição de genótipos de soja de maturação semiprecoce<sup>1</sup>.

| Modelo            | Estrutura <sup>2</sup> |      | Componente variância |                       |                    | Coef. det. genot. |               | Acurácia<br>( $\hat{r}_{gg}$ ) |
|-------------------|------------------------|------|----------------------|-----------------------|--------------------|-------------------|---------------|--------------------------------|
|                   | G                      | R    | $\hat{\sigma}_g^2$   | $\hat{\sigma}_{ga}^2$ | $\hat{\sigma}_e^2$ | $\hat{h}_i^2$     | $\hat{h}_m^2$ |                                |
| BLUE <sub>F</sub> | -                      | I    | -                    | -                     | 205287,5           | -                 | -             | -                              |
|                   | -                      | AR1  | -                    | -                     | 210481,7           | -                 | -             | -                              |
|                   | -                      | ARMA | -                    | -                     | 146344,6           | -                 | -             | -                              |
|                   | -                      | SP   | -                    | -                     | 210600,0           | -                 | -             | -                              |
|                   | -                      | UN   | -                    | -                     | 208439,9           | -                 | -             | -                              |
| BLUP <sub>E</sub> | I                      | I    | 12298,5              | 40249,8               | 206104,5           | 0,05              | 0,55          | 0,81                           |
|                   | I                      | AR1  | 9025,6               | 34557,7               | 217577,1           | 0,03              | 0,48          | 0,76                           |
|                   | I                      | ARMA | 6321,1               | 49618,4               | 146810,0           | 0,03              | 0,41          | 0,76                           |
|                   | I                      | SP   | 8353,4               | 35122,1               | 219199,9           | 0,03              | 0,46          | 0,74                           |
|                   | I                      | UN   | 8457,0               | 32777,7               | 217201,1           | 0,03              | 0,47          | 0,75                           |
| BLUP <sub>p</sub> | A <sub>G</sub>         | I    | 7881,1               | 19012,2               | 211752,4           | 0,03              | 0,49          | 0,74                           |
|                   | A <sub>G</sub>         | AR1  | 5701,5               | 17355,8               | 222051,0           | 0,02              | 0,41          | 0,67                           |
|                   | A <sub>G</sub>         | ARMA | 3451,2               | 41356,6               | 148651,6           | 0,02              | 0,29          | 0,66                           |
|                   | A <sub>G</sub>         | SP   | 5157,8               | 17728,4               | 223745,6           | 0,02              | 0,38          | 0,65                           |
|                   | A <sub>G</sub>         | UN   | 4955,3               | 15839,7               | 222273,3           | 0,02              | 0,38          | 0,65                           |
| BLUP <sub>M</sub> | A <sub>M</sub>         | I    | 7653,5               | 16840,5               | 213409,6           | 0,03              | 0,49          | 0,73                           |
|                   | A <sub>M</sub>         | AR1  | 5776,9               | 16267,6               | 222327,5           | 0,02              | 0,41          | 0,68                           |
|                   | A <sub>M</sub>         | ARMA | 3451,2               | 39421,0               | 148092,2           | 0,02              | 0,30          | 0,66                           |
|                   | A <sub>M</sub>         | SP   | 5210,4               | 16609,9               | 223411,7           | 0,02              | 0,39          | 0,66                           |
|                   | A <sub>M</sub>         | UN   | 5326,0               | 14852,5               | 222023,3           | 0,02              | 0,40          | 0,66                           |

<sup>1</sup> Onze experimentos conduzidos no delineamento de blocos completos casualizados, com 43 tratamentos e três repetições.

<sup>2</sup> Estruturas associadas à matriz G: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz R: identidade (I), autorregressiva de primeira ordem (AR1), autorregressiva com médias móveis (ARMA), exponencial geométrica (SP) e não estruturada (UN).

determinação genotípica ou herdabilidade, em nível de médias ( $\hat{h}_m^2$ ), que apresentaram magnitudes expressivamente superiores às obtidas em nível individual ( $\hat{h}_i^2$ ) (Tabelas 4.4 a 4.6). Especificamente para o grupo de genótipos superprecoces, esse coeficiente, em nível

de médias, foi de 0,93; valor que é relativamente elevado ao se tratar de um caráter quantitativo. Credita-se a esta estimativa a condição de elevado número de experimentos com bom controle da variação residual, associado à elevada variabilidade genética para esse caráter, nesse grupo de maturação.

Segundo alguns autores, a acurácia seletiva é o parâmetro estatístico de maior importância na avaliação genotípica (Henderson, 1984; Bernardo, 2002). Esse parâmetro refere-se à correlação entre o valor genotípico verdadeiro e aquele estimado ou predito a partir dos dados experimentais; assim, assume valores entre 0 e 1 (Bernardo, 2002; Resende & Duarte, 2007; Lorenzana & Bernardo, 2009), sendo considerados adequados aqueles próximos à unidade (Resende & Duarte, 2007). Embora não considerado no presente estudo, Souza Filho (2013) adverte sobre a necessidade de considerar, conjuntamente à acurácia, o tamanho amostral, representado pelo número de progênies e de repetições (ou parcelas) disponíveis para tal avaliação.

Conforme observado nas Tabelas 4.4 a 4.6, nos três grupos de experimentos a maior acurácia ( $\hat{r}_{gg}$ ) resultou do ajuste do modelo misto  $BLUP_E + I$ . Este assume apenas a interdependência genotípica associada à variância genética comum ( $\sigma_g^2$ ) utilizada na composição da matriz  $G$ , indicando que o uso da informação de parentesco não proporcionou melhoria na acurácia das predições. Resultado semelhante foi identificado por Piepho & Möhring (2007). Contrariamente, Pinheiro et al. (2013) observaram que a inclusão de parentesco aumentou o ganho em acurácia seletiva na proporção de 5,5% a 11,2%. Por isso, consideraram o seu uso adequado, principalmente em situações de menor precisão experimental. Nunes et al. (2008) também constataram que a inclusão de parentesco associado à matriz  $G$  resultou em maiores ganhos com a seleção de progênies de soja, além de estimativas de variância genética e de herdabilidade com menores erros de predição.

A inclusão de estruturas espaciais associadas às respectivas modelagens também resultou numa tendência de diminuição da acurácia seletiva. Apenas no grupo superprecoce, caracterizado por valores de  $\hat{h}_m^2$  elevados, a acurácia seletiva manteve-se com estimativas de elevada magnitude, variando entre 0,97 e 0,98 nos diferentes modelos (Tabela 4.4). Esta elevada acurácia, associada à pequena variação de suas estimativas nos diferentes modelos, está, naturalmente, relacionada à condição de  $\hat{h}_m^2$  também elevada. Nos

grupos precoce e semiprecoce, que apresentaram menores estimativas de  $h_m^2$ , foi possível identificar maior diferenciação de acurácia entre os modelos. Nestes dois casos, o modelo  $BLUP_E + I$ , com matrizes identidades associadas a  $G$  e  $R$ , foi responsável pelos maiores valores de acurácia, 0,79 e 0,81, (Tabelas 4.5 e 4.6, respectivamente).

Considerando a variância do erro de predição ( $PEV$ ), no grupo de genótipos superprecoces os modelos fixo e mistos não apresentaram grandes diferenças em suas estimativas, enquanto nos outros dois grupos (precoce e semiprecoce) essa variação foi expressiva (Tabela 4.7). Por essa estatística, o melhor ajuste é associado ao modelo com a menor estimativa de  $PEV$  (Resende & Duarte, 2007). Assim, no caso dos genótipos superprecoces, cuja herdabilidade média foi elevada (0,93), o modelo fixo não diferiu dos demais (modelos mistos), quanto ao erro médio de predição. Mas, nos grupos precoce e semiprecoce, cujas herdabilidades foram menores (0,52 e 0,55, respectivamente), os modelos mistos mostraram maior precisão nas suas predições. No primeiro grupo (superprecoce), os modelos com as menores estimativas de  $PEV$  foram aqueles com estrutura  $R = I\hat{\sigma}_e^2$ , sugerindo que sob herdabilidade alta os efeitos da falta de independência praticamente não afetam a discriminação estatística dos genótipos. Já sob herdabilidades mais baixas, ou apenas moderadas, a informação de parentesco ( $BLUP_P$  e  $BLUP_M$ ) mostra-se relevante para a melhoria da discriminação genotípica. Apesar disso, estas modelagens não foram as mais adequadas com base nos critérios  $AIC$  e  $BIC$ , nem com base na acurácia seletiva.

Em síntese, considerando-se os critérios de escolha dos modelos avaliados, baseados em estatísticas de informação de ajuste ( $AIC$  e  $BIC$ ), logREML, além da acurácia seletiva e erro médio de predição, associados à ocorrência de herdabilidades moderadas ou altas, o modelo fixo é o que apresenta melhor ajuste aos dados de produtividade de grãos objeto deste estudo. Mostrou-se, portanto, adequado e suficiente para a estimação dos valores genotípicos das linhagens de soja em avaliação. Vale reiterar que esta avaliação baseou-se numa estrutura experimental de delineamento em blocos completos casualizados, com uso de repetições.

Potencial melhoria na estimação desses efeitos genéticos também pode ser obtida removendo-se resíduos de autocorrelação espacial associadas à matriz  $R$ , principalmente usando-se a estrutura autorregressiva de primeira ordem (AR1), ou a estrutura exponencial geométrica (SP). Neste estudo, em que se observou autocorrelação

espacial significativa ( $p < 0,05$ ), embora de baixa magnitude (Apêndice A2), as modelagens espaciais proporcionaram melhorias na capacidade de ajustamento dos modelos, em comparação à tradicional estrutura de modelo fixo, que pressupõe independência dos erros ( $R = I\sigma_e^2$ ). Essa melhoria foi constatada também na avaliação da eficiência relativa (ER) dessas análises em comparação ao modelo fixo tradicional (Tabela 4.3), indicando que o ajuste espacial, sob tais condições, propicia maior eficiência na

**Tabela 4.7.** Estimativas da variância do erro de predição (PEV), obtidas com modelos de análise considerando diferentes informações associadas às matrizes de variâncias-covariâncias genéticas (G) e residuais (R), para dados de produtividade de grãos ( $\text{kg.ha}^{-1}$ ), em grupos de experimentos de competição de genótipos de soja, com distintos grupos de maturação e diferentes estimativas do coeficiente de determinação genotípica ( $\hat{h}_m^2$ ).

| Modelo                         | Estrutura <sup>1</sup> |     | Grupo de maturação                                    |                                     |                                         |
|--------------------------------|------------------------|-----|-------------------------------------------------------|-------------------------------------|-----------------------------------------|
|                                | G                      | R   | Superprecoce<br>( $\hat{h}_m^2 = 0,93$ ) <sup>2</sup> | Precoce<br>( $\hat{h}_m^2 = 0,52$ ) | Semiprecoce<br>( $\hat{h}_m^2 = 0,55$ ) |
| BLUE <sub>F</sub> <sup>3</sup> | -                      | I   | 2255,44                                               | 4699,64                             | 6318,77                                 |
|                                | -                      | AR1 | 2322,68                                               | 4303,36                             | 6478,65                                 |
|                                | -                      | SP  | 2334,89                                               | 4810,34                             | 6482,29                                 |
|                                | -                      | UN  | 2255,44                                               | 4687,20                             | 6415,80                                 |
| BLUP <sub>E</sub>              | I                      | I   | 2183,19                                               | 2903,04                             | 4185,11                                 |
|                                | I                      | AR1 | 2655,67                                               | 2861,99                             | 3844,45                                 |
|                                | I                      | SP  | 2402,15                                               | 2657,28                             | 3732,37                                 |
|                                | I                      | UN  | 2242,22                                               | 2676,92                             | 3733,81                                 |
| BLUP <sub>P</sub>              | A <sub>G</sub>         | I   | 2137,12                                               | 2324,44                             | 3567,45                                 |
|                                | A <sub>G</sub>         | AR1 | 2338,84                                               | 2363,24                             | 3108,44                                 |
|                                | A <sub>G</sub>         | SP  | 2354,67                                               | 2156,38                             | 2949,13                                 |
|                                | A <sub>G</sub>         | UN  | 2196,98                                               | 2185,67                             | 2873,83                                 |
| BLUP <sub>M</sub>              | A <sub>M</sub>         | I   | 2135,31                                               | 2202,59                             | 3534,88                                 |
|                                | A <sub>M</sub>         | AR1 | 2330,17                                               | 2269,78                             | 3132,51                                 |
|                                | A <sub>M</sub>         | SP  | 2340,46                                               | 2089,23                             | 2964,34                                 |
|                                | A <sub>M</sub>         | UN  | 2182,35                                               | 2128,17                             | 2993,21                                 |

<sup>1</sup> Estruturas associadas à matriz G: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz R: identidade (I), autorregressiva de primeira ordem (AR1), exponencial geométrica (SP) e não estruturada (UN).

<sup>2</sup> Coeficiente de determinação genotípica ou herdabilidade sentido amplo ( $\hat{h}_m^2$ ), em nível de médias, obtido do modelo de referência BLUP<sub>E</sub> + I (modelo misto, com pressuposição de erros independentes).

<sup>3</sup> As estimativas de PEV no modelo fixo (BLUE<sub>F</sub>) foram obtidas por:  $PEV = \hat{\sigma}_e^2 / ab$  (adaptado de Resende & Duarte, 2007); para este caso não valem as estimativas de herdabilidade apresentadas, haja vista a suposição de  $h_m^2 = 1,0$ .

avaliação experimental. Por isso, em condições nas quais a pressuposição de independência dos erros não é atendida, a estratégia de análise espacial garante a obtenção de estimativas mais realistas, assegurando maior acurácia e validade às inferências (Duarte & Vencovsky, 2005; Smith et al., 2005; Resende, 2007; Piepho & Williams, 2010; Leiser et al., 2012; Williams & Piepho, 2013).

A avaliação de modelos que flexibilizam a estrutura da matriz  $R$  com aplicação de técnicas de análise espacial, no contexto dos modelos mistos, representa, portanto, uma alternativa de controle da variação local sob condições de erros correlacionados (Duarte & Vencovsky, 2005; Piepho & Williams, 2010). Segundo Resende (2007), nem sempre a casualização e as formas de controle local baseadas em blocagem são suficientes para a neutralização dos efeitos de correlação espacial e, portanto, para a geração de uma análise de variância fidedigna. Nesse caso, essa é uma importante forma de controle e correção dos dados, sendo o seu uso recomendado, sempre que se observar a presença de autocorrelação espacial nos dados, o que se caracteriza pela não independência dos erros (Duarte & Vencovsky, 2005; Resende, 2007; Leiser et al., 2012; Piepho et al., 2013). Atualmente tem se observado maior popularização de delineamentos construídos para acomodar correlação espacial, em experimentos de competição de cultivares, e isso tem sido favorecido pela crescente disponibilidade de algoritmos computacionais que facilitam a geração e a análise dos delineamentos espaciais (Smith et al., 2005; Williams & Piepho, 2013). Deve-se ressaltar, entretanto, a preocupação de Smith et al. (2005), que advogam o uso da análise espacial não como substituto do modelo tradicional de experimentação, envolvendo blocagem e casualização, mas, sim, como complemento, visando a melhoria na análise por melhor acomodação de prováveis variações no campo experimental.

Ressalta-se também a atual disponibilidade de programas estatísticos e recursos computacionais que possibilitam a análise de modelos mais complexos, como a modelagem mista, com rapidez e eficiência. Dessa forma, é possível trabalhar com modelos teóricos que se aproximem cada vez mais da realidade, resultando em estimativas ou predições de maior acurácia e, conseqüentemente, em maior eficiência e qualidade no processo seletivo. Além disso, o uso de modelos mais generalistas permite melhor aproveitamento de experimentos e dados que, numa condição de análise tradicional, não raramente violam algumas de suas restritivas pressuposições, devendo, portanto, serem descartados. Um exemplo prático disso seria a exclusão de experimentos que não apresentassem homogeneidade de variâncias (Apêndice A2). Logo, por não se adequarem

à análise tradicional de grupos de experimentos, não poderiam ser aproveitados. Contrariamente, no presente estudo todos os ensaios foram considerados no apoio às inferências.

#### **4.3.3 Efeito da redução de repetições e do desbalanceamento na escolha do modelo**

Os resultados obtidos com os exercícios de simulação de desbalanceamento nos dados ou de redução do número de repetições indicaram que, com referência aos três grupos de maturação, o modelo fixo continuou apresentando os menores valores estimados para o critério de informação de Akaike (*AIC*). Assim, revelou-se a alternativa de melhor ajuste aos dados, mesmo nas condições simuladas com menos repetições ou com diferentes níveis de desbalanceamento (Tabela 4.8). Provavelmente, a robustez identificada na análise de modelo fixo, mesmo nessas condições, deve-se à adequação dos experimentos avaliados no que concerne ao atendimento das pressuposições da análise tradicional de dados; principalmente, no que se refere aos cuidados no planejamento e na definição de repetições, casualização e controle local, adequados.

Por fim, deve-se ressaltar que, embora boa parte dos resultados do presente estudo não apoie a adoção irrestrita da abordagem de modelos mistos, haja vista a identificação de modelagem mais parcimoniosa e acurada, isto não representa, de modo algum, descrédito às suas potencialidades. Pelo contrário, estimula-se a continuidade de sua aplicação e avaliação permanente, haja vista suas características de generalidade e de eficiência estatística, esta última também aqui demonstrada, que permitem aproveitamento otimizado de informações disponíveis em programas de melhoramento, e que não podem ser exploradas conjuntamente num contexto de análise tradicional de modelo fixo.

**Tabela 4.8.** Critério de informação de Akaike (AIC) estimado a partir de diferentes modelos de análise, considerando-se redução do número de repetições e níveis distintos de desbalanceamento provocado nos dados, para produtividade de grãos (kg.ha<sup>-1</sup>), num conjunto de experimentos de competição de genótipos de soja de três grupos de maturação.

| Modelo            | Estrutura <sup>1</sup> |      | Grupo Superprecoce |              |              |              |              | Grupo Precoce |              |              |              |              | Grupo Semiprecoce |              |              |              |              |
|-------------------|------------------------|------|--------------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|-------------------|--------------|--------------|--------------|--------------|
|                   | G                      | R    | Np <sup>2</sup>    | DBC/<br>3rep | DBC/<br>2rep | DBC/<br>Unb1 | DBC/<br>Unb2 | Np            | DBC/<br>3rep | DBC/<br>2rep | DBC/<br>Unb1 | DBC/<br>Unb2 | Np                | DBC/<br>3rep | DBC/<br>2rep | DBC/<br>Unb1 | DBC/<br>Unb2 |
| BLUE <sub>F</sub> | -                      | I    | 1                  | 16927,9      | 8588,8       | 9342,1       | 9119,0       | 1             | 14324,4      | 7336,5       | 7921,2       | 4407,7       | 1                 | 12517,9      | 6191,6       | 6707,8       | 3343,9       |
|                   | -                      | AR1  | 69                 | 16808,7      | 8555,2       | 9308,0       | 9066,0       | 42            | 14155,4      | 7261,0       | 7839,7       | 4365,0       | 33                | 12346,5      | 6132,7       | 6630,4       | 3329,0       |
|                   | -                      | ARMA | 115                | 17066,5      | 8769,7       | 9512,7       | 9290,8       | 70            | 14388,3      | 7438,3       | 8015,2       | 4531,0       | 55                | 12545,3      | 6250,5       | 6768,9       | 3447,1       |
|                   | -                      | SP   | 46                 | 16792,2      | 8544,9       | 9298,2       | 9068,9       | 28            | 14160,8      | 7253,7       | 7831,8       | 4367,1       | 22                | 12340,5      | 6125,7       | 6616,5       | 3315,5       |
|                   | -                      | UN   | 23                 | 16820,7      | 8546,2       | 9297,2       | 9071,7       | 14            | 14165,3      | 7251,9       | 7819,0       | 4370,8       | 11                | 12362,0      | 6125,6       | 6619,1       | 3306,5       |
| BLUP <sub>E</sub> | I                      | I    | 5                  | 25755,8      | 17241,7      | 18127,9      | 18102,7      | 5             | 21769,4      | 14772,2      | 15244,8      | 11257,0      | 5                 | 18832,6      | 12527,8      | 12908,5      | 8799,4       |
|                   | I                      | AR1  | 73                 | 25563,5      | 17139,3      | 18036,7      | 17953,6      | 46            | 21584,9      | 14655,0      | 15126,4      | 11208,7      | 37                | 18582,7      | 12402,4      | 12746,8      | 8718,2       |
|                   | I                      | ARMA | 119                | 25989,2      | 17494,2      | 18386,7      | 18338,2      | 74            | 21947,2      | 14958,0      | 15445,3      | 11467,1      | 59                | 18941,3      | 12651,9      | 13048,0      | 8966,5       |
|                   | I                      | SP   | 50                 | 25548,1      | 17123,5      | 18020,4      | 17953,1      | 32            | 21590,6      | 14653,6      | 15123,0      | 11195,3      | 26                | 18581,4      | 12397,2      | 12738,9      | 8708,7       |
|                   | I                      | UN   | 27                 | 25620,1      | 17169,7      | 18044,0      | 18008,4      | 18            | 21595,1      | 14648,8      | 15120,2      | 11187,8      | 15                | 18627,9      | 12407,0      | 12766,8      | 8709,8       |
| BLUP <sub>P</sub> | A <sub>G</sub>         | I    | 5                  | 25732,0      | 17225,9      | 18124,4      | 18085,2      | 5             | 21738,6      | 14752,6      | 15233,3      | 11238,7      | 5                 | 18842,9      | 12535,6      | 12915,8      | 8797,7       |
|                   | A <sub>G</sub>         | AR1  | 73                 | 25539,7      | 17127,8      | 18024,0      | 17944,4      | 46            | 21544,2      | 14620,6      | 15114,3      | 11186,6      | 37                | 18588,8      | 12409,4      | 12754,3      | 8718,5       |
|                   | A <sub>G</sub>         | ARMA | 119                | 25932,1      | 17454,7      | 18352,7      | 18303,3      | 74            | 21914,0      | 14924,7      | 15437,7      | 11445,5      | 59                | 18986,6      | 12680,2      | 13082,9      | 8986,2       |
|                   | A <sub>G</sub>         | SP   | 50                 | 25522,3      | 17108,0      | 18007,6      | 17943,6      | 32            | 21549,4      | 14619,6      | 15109,4      | 11174,2      | 26                | 18587,9      | 12404,5      | 12746,6      | 8707,7       |
|                   | A <sub>G</sub>         | UN   | 27                 | 25588,6      | 17148,8      | 18029,6      | 17987,9      | 18            | 21557,6      | 14618,3      | 15106,0      | 11168,7      | 15                | 18635,6      | 12414,6      | 12773,6      | 8708,1       |
| BLUP <sub>M</sub> | A <sub>M</sub>         | I    | 5                  | 25727,9      | 17224,5      | 18117,2      | 18090,4      | 5             | 21738,8      | 14748,7      | 15233,6      | 11245,6      | 5                 | 18842,7      | 12540,5      | 12912,3      | 8799,9       |
|                   | A <sub>M</sub>         | AR1  | 73                 | 25532,5      | 17129,1      | 18014,3      | 17949,9      | 46            | 21551,0      | 14624,8      | 15114,5      | 11193,0      | 37                | 18586,4      | 12411,2      | 12749,1      | 8719,0       |
|                   | A <sub>M</sub>         | ARMA | 119                | 25944,7      | 17469,3      | 18359,4      | 18322,9      | 74            | 21941,5      | 14947,3      | 15460,6      | 11471,1      | 59                | 18998,3      | 12705,9      | 13090,8      | 9006,4       |
|                   | A <sub>M</sub>         | SP   | 50                 | 25519,7      | 17112,9      | 18001,6      | 17948,3      | 32            | 21554,5      | 14621,5      | 15110,6      | 11178,8      | 26                | 18587,0      | 12408,5      | 12742,4      | 8708,2       |
|                   | A <sub>M</sub>         | UN   | 27                 | 25586,6      | 17150,4      | 18023,0      | 17995,3      | 18            | 21563,2      | 14619,0      | 15108,8      | 11174,6      | 15                | 18634,3      | 12418,7      | 12771,1      | 8708,8       |

<sup>1</sup> Estruturas associadas à matriz G: matriz identidade (I) e matrizes de parentesco obtidas por genealogia (A<sub>G</sub>) e por marcadores SSR (A<sub>M</sub>); e, à matriz R: identidade (I), autorregressiva de primeira ordem (AR1), autorregressiva com médias móveis (ARMA), exponencial geométrica (SP) e não estruturada (UN). <sup>2</sup> Np, DBC/3rep, DBC/2rep, DBC/Unb1 e DBC/Unb2 indicam, respectivamente, número de parâmetros, blocos completos casualizados com 3 repetições, blocos completos casualizados com duas repetições, blocos casualizados com 30% de desbalanceamento e blocos casualizados com 50% de desbalanceamento.



#### 4.4 CONCLUSÕES

- i) Corrobora-se a tese de que a aplicação de modelo fixo é adequada para estimar efeitos genotípicos, sob condições de balanceamento dos dados e ortogonalidade do delineamento, normalmente atendidas em ensaios de competição final de genótipos de soja.
- ii) Nessas condições e sob herdabilidade moderada ou baixa, a inclusão da informação de parentesco, advinda de genealogias ou de marcadores moleculares, proporciona aumento na precisão das predições de valores genotípicos.
- iii) O uso de estruturas espaciais associadas ao modelo de análise, na presença de autocorrelação residual, melhora o ajuste dos modelos, mesmo numa condição de balanceamento dos dados e casualização, podendo favorecer a seleção em ensaios de competição de genótipos.
- iv) A escolha do modelo de análise, em condições de altos coeficientes de herdabilidade, não altera o posicionamento relativo dos genótipos; logo, não impacta significativamente na seleção de genótipos superiores.

## 5 CONCLUSÕES GERAIS

A iniciativa de estudos que favoreçam a identificação de germoplasma de alto valor para uso em planos de hibridação, bem como na identificação de métodos de estimação ou predição de valores genotípicos de maior eficiência e acurácia, é de grande relevância para suporte aos programas de melhoramento de plantas. Sobretudo em espécies já bastante melhoradas como a soja, em que os ganhos genéticos tornam-se cada vez mais difíceis, mesmo numa condição de alta capacidade de intercâmbio (introduções), desenvolvimento e teste de germoplasma apresentados pelos programas de pesquisa nos dias atuais.

Nesse enfoque, considerando os resultados principais deste estudo, pode-se verificar que os marcadores microssatélites são adequados ao estudo de divergência genética, mostrando-se consistentes na identificação de grupos genotípicos divergentes e de subpopulações em soja, possibilitando, assim, o planejamento de combinações híbridas com maior variabilidade genética. Também constatou-se que a introdução de germoplasma melhorado garante a incorporação de alelos necessários à ampliação da base genética da soja cultivada nos trópicos brasileiros, e, conseqüentemente, da respectiva variabilidade para uso no processo seletivo. Com isso, permite-se a obtenção de populações base adequadas à seleção, favorecendo a continuidade dos ganhos genéticos com o melhoramento nessa espécie.

Com respeito à análise de ensaios de competição final de genótipos, evidenciou-se que o modelo fixo, em experimentos com balanceamento dos dados e ortogonalidade do delineamento, propicia melhor ajuste aos dados em relação ao modelo aleatório ou misto, independentemente do uso de informação de parentesco associada à matriz  $G$  (covariância genética) e, ou, estrutura espacial associada à matriz  $R$  (covariância residual). Adicionalmente, sob as condições desta pesquisa, constatou-se que o ordenamento de genótipos por seus valores genotípicos apresenta alta correlação entre os distintos modelos, indicando não haver alterações nas estimativas ou predições de valores genotípicos, quando usadas diferentes modelagens. Portanto, as diferentes modelagens não impactaram na identificação e seleção de genótipos superiores. Estes resultados reforçam a

tese de eficiência e robustez da análise tradicional de modelos fixos na estimação de valores genotípicos sob condição de balanceamento dos dados e uso de repetições. Logo, o uso de modelos parcimoniosos, a exemplo do modelo fixo associado à análise de variância clássica, favorece a execução da análise biométrica no contexto de ensaios de competição de genótipos de soja. Por outro lado, não se podem descartar as características vantajosas da abordagem de modelos mistos associada ao uso de informação de parentesco, sobretudo em caracteres de moderada ou baixa herdabilidade, no que tange à obtenção de predições genotípicas de maior precisão.

## 6 REFERÊNCIAS

ANDRÉ, C. M. G. **Avaliação da melhor predição linear não tendenciosa (BLUP) associada ao uso de marcadores moleculares na análise dialélica**. 1999. 101 f. Dissertação (Mestrado em Agronomia: Genética e Melhoramento de Plantas)– Universidade Federal de Lavras, Lavras, 1999.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, New York, v. 19, n. 6, p. 716-723, 1974.

ARANTES, N. E.; KIIHL, R. A. S.; ALMEIDA, L. A.; ZITO, R. K.; YORINORI, J. T.; DIAS, W. P.; SOUZA, P. I. M.; NUNES FILHO, J. Cultivar de soja BRS Valiosa RR. In: REUNIÃO DE PESQUISA DE SOJA DA REGIÃO CENTRAL DO BRASIL, 27., 2005, Cornélio Procopio. **Resumos...** Londrina: Embrapa soja, 2005. p. 394-395.

BAUER, A. M.; REETZ, T. C.; LÉON, J. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. **Crop Science**, Madison, v. 46, n. 6, p. 2685-2691, 2006.

BERNARDO, R. Prediction of maize single-cross performance using RFLP and information from related hybrids. **Crop Science**, Madison, v. 34, p. 20-15, 1994.

BERNARDO, R. **Breeding for quantitative traits in plants**. Stemma Press, Woodbury, 2002. 359 p.

BERNARDO, R.; YU, J. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, Madison, v. 47, n. 3, p. 1082-1090, 2007.

BUENO FILHO, J. S. S. **Modelos mistos na predição de valores genéticos aditivos em testes de progênies florestais**. 1997. 118 f. Tese (Doutorado)–Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1997.

BUTLER, D. G. **Asreml**: asreml() fits the linear mixed model. R package version 3.0. 2009. Disponível em: <<http://www.vsni.co.uk/downloads/asreml>>. Acesso em: 29 ago. 2013.

BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J. **ASReml-R reference manual**: ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML). 3. ed. Queensland: Queensland Department of Primary Industries and Fisheries, 2009. 148 p. Disponível em: <<http://www.vsni.co.uk/downloads/asreml/release2/doc/asreml-R.pdf>>. Acesso em: 28 ago. 2013

CAMARINHA FILHO, J. A. **Modelos lineares mistos**: estruturas de matrizes de variâncias e covariâncias e seleção de modelos. 2002. 85 f. Tese (Doutorado em

Agronomia: Estatística e Experimentação Agronômica)–Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2002.

CARNEIRO, P. L. S.; MALHADO, C. H. M.; EUCLYDES, R. F.; TORRES, R. A.; LOPES, P. S.; CARNEIRO, A. P. S.; CUNHA, E. E. Seleção tradicional e associada a marcadores moleculares na avaliação genética animal. **Pesquisa Agropecuária Brasileira**, Brasília, v. 41, n. 4, p. 615-621, 2006.

COELHO, A. S. G. **Bood**: avaliação de dendrogramas baseados em estimativas de distâncias/similaridades genéticas através do procedimento de bootstrap. Goiânia: UFG. 2000.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira de grãos – safra 2013/14**: sétimo levantamento, abril 2014. Brasília: Conab, 2014. 92 p. Disponível em: <[http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14\\_04\\_11\\_16\\_42\\_24\\_boletim\\_graos\\_abril\\_2014.pdf](http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_04_11_16_42_24_boletim_graos_abril_2014.pdf)>. Acesso em: 10 abr. 2014.

CREGAN, P. B.; JARVIK, T.; BUSH, A. L.; SHOEMAKER, R. C.; LARK, K. G.; KAHLER, A. L.; KAYA, N.; VANTOAI, T. T.; LOHNES, D. G.; CHUNG, J.; SPECHT, J. E. An integrated genetic linkage map of the soybean genome. **Crop Science**, Madison, v. 39, n. 5, p. 1464-1490, 1999.

DEMPSTER, A.; LAIR, N.; RUBIN, D. Maximum likelihood from incomplete data via the EM Algorithm. **Journal of the Royal Statistical Society Series B Statistical Methodology**, London, v. 39, n. 1, p. 1-38, 1977.

DOYLE, J. J.; DOYLE, J. L. Isolation of DNA from fresh tissue. **Focus**, Rockville, v. 12, n. 27, p. 13-15, 1990.

DUARTE, J. B. **Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal**. 2000. 293 f. Tese (Doutorado em Agronomia: Genética e Melhoramento de Plantas)–Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2000.

DUARTE, J. B. Biometria em genética e melhoramento de plantas: tendências e inquietações. In: SIMPÓSIO DE ATUALIZAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS, 14., 2010. Lavras. **Anais...** Lavras: UFLA, 2010. 1 CD-ROM.

DUARTE, J. B.; VENCOSKY, R. Estimação 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.

DUARTE, J. B.; VENCOSKY, R. Spatial statistical analysis and selection of genotypes in plant breeding. **Pesquisa Agropecuária Brasileira**, Brasília, v. 40, n. 2, p. 107-114, 2005.

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

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software Structure: a simulation study. **Molecular Ecology Notes**, Oxford, v. 14, n. 8, p. 2611-2620, 2005.

FALCONER, D. S.; MACKAY, T. F. C. **An introduction to quantitative genetics**. 4. ed. London: Prentice Hall, 1996. 480 p.

FELIPE, C. R. P.; DUARTE, J. B. Comparison of estimation and prediction methods of genotypic means in maize variety trials. **Crop Breeding and Applied Biotechnology**, Londrina, v. 10, n. 2, p. 147-153, 2010.

FERRAROTTI, J. Soybean breeding program in Argentina. In: WORLD SOYBEAN RESEARCH CONFERENCE, 7., 2004, Foz do Iguaçu. **Annals...** Londrina: Embrapa soybean, 2004. p. 191-201.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3. ed. Brasília: Embrapa Cenargen, 1998. 220 p.

GARCIA, A. F.; ALBERINI, J. L.; ZUCCHI, M. I.; SOUZA, A. P. Microsatellite molecular markers in the cultivar identification of Brazilian soybean for human consumption. **Crop Breeding and Applied Biotechnology**, Londrina, v. 7, n. 2, p. 155-164, 2007.

GAZZONI, D. L. Botánica. In: EMBRAPA-CNPSo (Ed.). **El cultivo de la soja en los tropicos: mejoramiento y producción**. Roma: FAO, 1995. p. 1-12. (Colección FAO, n. 15)

GIZLICE, Z.; CARTER, T. E.; BURTON, J. W. Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. **Crop Science**, Madison, v. 34, n. 5, p. 1143-1151, 1994.

GUO, J.; WANG, Y.; SONG, C.; ZHOU, J.; QIU, L.; HUANG, H.; WANG, Y. A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. **Annals of Botany**, Oxford, v. 106, n. 3, p. 505-514, 2010.

HARDY, O. J., VEKEMANS, X. Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. **Molecular Ecology Notes**, Oxford, v. 2, n. 5, p. 618-620, 2002.

HENDERSON, C. R. **Applications of Linear Models in Animal Breeding**. Guelph: University of Guelph, 1984. 462 p.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; vonKROSIGK, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, Washington, v. 13, n. 1, p. 192-218, 1959.

HIROMOTO, D. M.; VELLO, N. A. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazilian Journal of Genetics**, Ribeirão Preto, v. 9, n. 2, p. 295-306, 1986.

HWANG, T.; SAYAMA, T.; TAKAHASHI, M.; TAKADA, Y.; NAKAMOTO, Y.; UNATSUKI, H.; HISANO, H.; SASAMOTO, S.; SATO, S.; TABATA, S.; KONO, I.; HOSHI, M.; HANAWA, M.; YANO, C.; XIA, Z.; HARADA, K.; KITAMURA, K.; ISHIMOTO, M. High-density Integrated linkage Map Based on SSR Markers in Soybean. **DNA Research**, Oxford, v. 16, n. 4, p. 213-225, 2009.

HYMOWITZ, T. Speciation and cytogenetics. In: BOERMA, H. J.; SPECHT, J. E. (Ed.). **Soybeans: improvement, production and uses**. 3. ed. Madison: American Society of Agronomy, 2004. p. 97-136.

HYTEN, D. L.; SONG, Q.; ZHU, Y.; CHOI, I.; NELSON, R. L.; COSTA, J. M.; SPECHT, J. E.; SHOEMAKER, R. C.; CREGAN, P. B. Impacts of genetic bottlenecks on soybean genome diversity. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 103, n. 45, p. 16666-16671, 2006.

KASTER, M.; FARIAS, J. R. B. **Regionalização dos testes de Valor de Cultivo e Uso e da indicação de cultivares de soja – terceira aproximação**. Londrina: Embrapa Soja, 2012. p. 69. Disponível em: <[http://www.cnpso.embrapa.br/download/Doc\\_330-OL.pdf](http://www.cnpso.embrapa.br/download/Doc_330-OL.pdf)>. Acesso em: 08 abr. 2014.

KELLY, A. M.; CULLIS, B. R.; GILMOUR, A. R.; ECCLESTON, J. A.; THOMPSON, R. Estimation in a multiplicative mixed model involving a genetic relationship matrix. **Genetics Selection Evolution**, Paris, v. 41, n. 4, p. 1-9, 2009.

KEMPTHORNE, O. **An introduction to genetics statistics**. New York: John Wiley, 1957. 545 p.

KIIHL, R. A. S. Soja: realidade e tendências. In: REUNIÃO DE PESQUISA DE SOJA DA REGIÃO SUL, 39., 2012, Passo Fundo. **Anais...** Passo Fundo: Embrapa trigo, 2012. Disponível em: <[http://www.cnpt.embrapa.br/39rpsojars/recursos/palestras/seminario/tendencia/realidade\\_tendencias\\_romeu\\_kiihl.pdf](http://www.cnpt.embrapa.br/39rpsojars/recursos/palestras/seminario/tendencia/realidade_tendencias_romeu_kiihl.pdf)>. Acesso em: 15 maio 2013.

KLEFFMANNGROUP. **Agricultural marketing information system (AMIS): seeds soybean 12/13**. 2013. Valinhos: Kleffmanngroup, 2013. 83 p.

LEISER, W. L.; RATTUNDE, H. F.; PEPHO, H. P.; PARZIES, H. K. Getting the most out of sorghum low-input field trials in west Africa using spatial adjustment. **Journal of Agronomy and Crop Science**, Berlin, v. 198, n. 5, p. 349-359, 2012.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D. **SAS system for mixed models**. Cary: SAS Institute Inc., 1996. 633 p.

LORENZANA, R. E.; BERNARDO, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and Applied Genetics**, New York, v. 120, n. 1, p. 151-161, 2009.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, p. 1819-1829, 2001.

MIKEL, M. A.; DIERS, B. W.; NELSON, R. L.; SMITH, H. H. Genetic diversity and agronomic improvement of North American soybean germplasm. **Crop Science**, Madison, v. 50, n. 4, p. 1219-1229, 2010.

MILLER, M. **Tools for population genetic analyses (TFPGA)**: a windows program for the analysis of allozyme and molecular population genetic data, version 1.3. Flagstaff: Northern Arizona University, 1997. 33 p. Disponível em: <<http://www.marksgeneticsoftware.net/tfpga.htm>>. Acesso em: 28 jun. 2012.

MIRANDA, Z. F. S. **Base genética de cultivares de soja no Brasil**. 2005. 871 f. Tese (Doutorado em Genética e Melhoramento de Plantas)—Universidade Estadual de Londrina, Londrina, 2005.

MIRANDA, Z. F. S.; ARIAS, C. A. A.; PRETE, C. E. C.; KIIHL, R. A. S.; ALMEIDA, L. A.; TOLEDO, J. F. F.; DESTRO, D. Genetic characterization of ninety elite soybean cultivars using coefficient of parentage. **Pesquisa Agropecuária Brasileira**, Brasília, v. 42, n. 3, p. 363-369, mar. 2007.

MORGANTE, M.; HANAFEY, M.; POWELL, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. **Nature Genetics**, New York, v. 30, n. 2, p. 194-200, 2002.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. Biddles: Guildford, 1996. 184 p.

MULATO, B. M.; MÖLLER, M.; ZUCCHI, M. I.; QUENCINI, V.; PINHEIRO, J. B. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 3, p. 276-283, 2010.

NEI, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences**, San Diego, v. 70, n. 12, p. 3321-3323, 1973.

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.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, London, v. 58, p. 545-554, 1971.

PIEPHO, H. P.; MÖHRING, J. Computing heritability and selection response from unbalanced plant breeding trials. **Genetics**, New York, v. 177, n. 3, p. 1881-1888, 2007.

PIEPHO, H. P.; WILLIAMS, E. R. Linear variance models for plant breeding trials. **Plant Breeding**, Berlin, v. 129, n. 1, p. 1-8, 2010.



PIEPHO, H. P.; MÖHRING, J.; MELCHINGER, A. B. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, Dordrecht, v. 161, p. 209-228, 2008.

PIEPHO, H. P.; MÖHRING, J.; WILLIAMS, E. R. Why Randomize Agricultural Experiments? **Journal of Agronomy and Crop Science**, Berlin, v. 199, n. 5, p. 374-383, 2013.

PINHEIRO, L. C. M.; GOOD GOD, P. I. V.; FARIA, V. R.; OLIVEIRA, A. G.; HASUI, A. A.; PINTO, E. H. G.; ARRUDA, K. M. A.; PIOVESAN, N. D.; MOREIRA, M. A. 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.

PRIOLLI, R. H. G.; MENDES JR., C. T.; ARANTES, N. E.; CONTEL, E. P. B. Characterization of Brazilian soybean cultivar using microsatellite markers. **Genetics and Molecular Biology**, Ribeirão Preto, v. 25, n. 2, p. 185-193, 2002.

PRIOLLI, R. H. G.; MENDES JR., C. T.; SOUZA, S. M. B.; ARANTES, N. E.; CONTEL, E. P. B. Diversidade genética da soja entre períodos e entre programas de melhoramento no Brasil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 39, n. 10, p. 967-975, 2004.

PRIOLLI, R. H. G.; PINHEIRO, J. B.; ZUCCHI, M. I.; BAIÁV, M. M.; VELLO, N. A. Genetic diversity among Brazilian soybean cultivars based on SSR loci and pedigree data. **Brazilian Archives of Biology and Technology**, Curitiba, v. 53, n. 3, p. 519-531, 2010.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Austin, v. 155, n. 2, p. 945-959, 2000.

R CORE TEAM (2013). **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2013. Vienna, Austria. Disponível em: <<http://www.R-project.org/>>. Acesso em: 28 ago. 2013.

RAMALHO, M. A. P.; ABREU, A. F. B.; SANTOS, J. B. Melhoramento de espécies autógamas. In: NASS, L. L.; VALOIS, A. C. C.; MELO, I. S.; VALADARES-IGLIS, M. E. (Ed.). **Recursos genéticos e melhoramento: plantas**. Rondonópolis: Fundação Mato Grosso, 2001. p. 201-230.

RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2000. 326 p.

REIF, J. C.; MELCHINGER, A. E.; FRISCH, M. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. **Crop Science**, Madison, v. 45, n. 1, p. 1-7, 2005.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 2002. 975 p.

RESENDE, M. D. V. **Métodos estatísticos ótimos na análise de experimentos de campo**. Colombo: Embrapa Florestas, 2004. 57 p.

RESENDE, M. D. V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 362 p.

RESENDE, M. D. V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, Goiânia, v. 37, n. 3, p. 182-194, 2007.

RITLAND, K. Estimators for pairwise relatedness and individual inbreeding coefficients. **Genetical Research**, London, v. 67, n. 2, p. 175-185, 1996.

ROHLF, F. J. **NTSYSpc**: numerical taxonomy and multivariate analysis system, version 2.1. Setauket: Exeter Software, 2000. 38 p.

ROSSI, R. L. Current status of the soybean production and utilization in Argentina. In: WORLD SOYBEAN RESEARCH CONFERENCE, 7., 2004, Foz do Iguaçu. **Annals...** Londrina: Embrapa soybean, 2004. p. 38-49.

ROSSI, R. L.; SARTORI, S. Genetics and physiology aspects of the indeterminate and determinate soybeans *Glycine max* (L.) Merrill and its relation with earliness. In: CONGRESSO BRASILEIRO DE SOJA, 5., 2009, Goiânia. **Anais...** Londrina: Embrapa soja, 2009. 1 CD-ROM.

SAS INSTITUTE INC. **SAS/STAT® 9.2 user's guide**. Cary: SAS Institute Inc., 2008. 7857 p. Disponível em: <<http://support.sas.com/documentation/cd1/en/statug/59654/PDF/default/statug.pdf>>. Acesso em: 16 mai. 2011.

SAYAMA, T.; HWANG, T.; KOMATSU, K.; TAKADA, Y.; TAKAHASHI, M.; KATO, S.; SASAMA, H.; HIGASHI, A.; NAKAMOTO, Y.; FUNATSUKI, H.; ISHIMOTO, M. Development and Application of a Whole-Genome Simple Sequence Repeat Panel for High-Throughput Genotyping in Soybean. **DNA Research**, New York, v. 18, n. 2, p. 107-115, 2011.

SCHMUTZ, J.; CANNON, S. B.; SCHLUETER, J.; MA, J.; MITROS, T.; NELSON W.; HYTEN, D. L.; SONG, Q.; THELEN, J. J.; CHENG, J.; XU, D.; HELLSTEN, U.; MAY, G. D.; YU, Y.; SAKURAI, T.; UMEZAWA, T.; BHATTACHARYYA, M. K.; SANDHU, D.; VALLIYODAN, B.; LINDQUIST, E.; PETO, M.; GRANT, D.; SHU, S.; GOODSTEIN, D.; BARRY, K.; FUTRELL-GRIGGS, M.; ABERNATHY, B.; DU, J.; TIAN, Z.; ZHU, L.; GILL, N.; JOSHI, T.; LIBAULT, M.; SETHURAMAN, A.; ZHANG, X. C.; SHINOZAKI, K.; NGUYEN, H. T.; WING, R. A.; CREGAN, P.; SPECHT, J.; GRIMWOOD, J.; ROKHSAR, D.; STACEY, G.; SHOEMAKER, R. C.; JACKSON, S. A. Genome sequence of the palaeopolyploid soybean. **Nature**, London, v. 463, n. 120, p. 178-183, 2010.

SCHWARZ, G. E. Estimating the dimension of a model. **Annals of Statistics**, Cleveland, v. 6, n. 2, p. 461-464, 1978.

SEARLE, S. R. **Matrix algebra useful for statistics**. New York: John Wiley & Sons, 1982. 438 p.

SEARLE, S. R.; CASELLA, G.; McCULLOCH, C. E. **Variance components**. New York: John Wiley & Sons, 1992. 501 p.

SEDIYAMA, T.; TEIXEIRA, R. C.; REIS, M. S. Melhoramento da soja. In: BORÉM, A. (Ed.). **Melhoramento de espécies cultivadas**. Viçosa: Editora da UFV, 1999. p. 487-533.

SILVA FILHO, J. L. **Estratégias de predição de cruzamentos não realizados em análise dialélica, utilizando informações de marcadores moleculares**. 2004. 70 f. Tese (Doutorado em Agronomia: Genética e Melhoramento de Plantas)–Universidade Federal de Lavras, Lavras, 2004.

SILVA FILHO, J. L. Optimizing the number of progenies and replications in plant breeding experiments. **Crop Breeding and Applied Biotechnology**, Londrina, v. 13, n. 3, p. 151-157, 2013.

SINGH, R. K.; BHATIA, V. S.; BHAT, K. V.; MOHAPATRA, T.; SING, N. K.; BANSAL, K. C.; KOUNDAL, K. R. SSR and AFLP based genetic diversity of soybean germoplasm differing in photoperiod sensitivity. **Genetics and Molecular Biology**, Ribeirão Preto, v. 33, n. 2, p. 319-324, 2010.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials : an overview of current mixed model approaches. **Journal of Agricultural Science**, London, v. 143, p. 449-462, 2005.

SOYBASE. **SoyBase and the soybean breeder's toolbox**: integrating genetics and molecular biology for soybean researchers. Disponível em: <<http://soybase.org/>>. Acesso em: 19 fev. 2013.

SOUZA, A. P. Biologia molecular aplicada ao melhoramento. In: NASS, L. L. **Recursos genéticos vegetais** (Ed.). Brasília: Embrapa, 2007. p. 939-965.

TOLEDO, J. F. F.; ALMEIDA, L. A.; KIIHL, R. A.; MENOSSO, O. G. Ganho genético em soja no Estado do Paraná via melhoramento. **Pesquisa Agropecuária Brasileira**, Brasília, v. 25, n. 1, p. 89-94, 1990.

TOLEDO, J. F. F.; ALMEIDA, L. A.; KIIHL, R. A. S.; KASTER, M.; ARIAS, C. A. A.; PIPOLO, A. E.; CARNEIRO, G. E. S. Soybean genetic breeding in Brazil. In: WORLD SOYBEAN RESEARCH CONFERENCE, 7., 2004, Foz do Iguaçu. **Annals...** Londrina: Embrapa soybean, 2004. p. 209-215.

VELLO, N. A.; HIROMOTO, D. M.; AZEVEDO FILHO, A. J. B. V. Coefficient of parentage and breeding of Brazilian soybean germplasm. **Brazilian Journal of Genetics**, Ribeirão Preto, v. 11, n. 3, p. 679-697, 1988.

VIEIRA, E. S. N.; SCHUSTER, I.; SILVA, R. B.; OLIVEIRA, M. A. R. Variabilidade genética em cultivares de soja determinada com marcadores microssatélites em gel de agarose. **Pesquisa Agropecuária Brasileira**, Brasília, v. 44, n. 11, p. 1460-1466, 2009.

YAMANAKA, N.; SATO, H.; YANG, Z.; XU, D. H.; CATELLI, L. L.; BINNECK, E.; ARIAS, C. A. A.; ABDELNOOR, R. V.; NEPOMUCENO, A. L. Genetic relationships between Chinese, Japanese and Brazilian soybean gene pools revealed by simple sequence repeat (SSR) markers. **Genetics and Molecular Biology**, Ribeirão Preto, v. 30, n. 1, p. 85-88, 2007.

WANG, K.; LI, X. Genetic differentiation and diversity of phenotypic characters in Chinese wild soybean (*Glycine soja* Sieb. et Zucc.) revealed by nuclear SSR markers and the implication for intraspecies phylogenetic relationship of characters. **Genetic Resources and Crop Evolution**, Dordrecht, v. 58, p. 209-223, 2011.

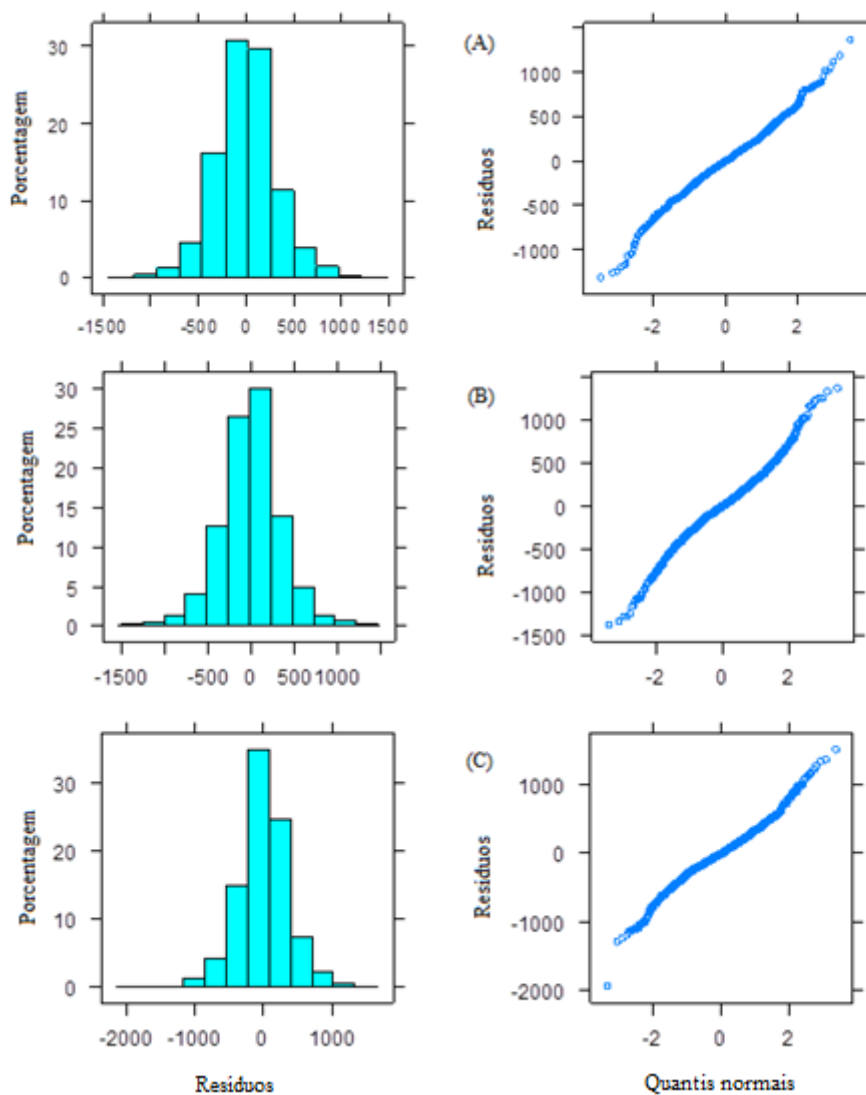
WEIR, B. S. **Genetic data analysis II**. 2. ed. Sunderland: Sinauer Association, 1996. 445 p.

WHITE, T. L.; HODGE, G. R. **Predicting breeding values with applications in forest tree improvement**. Dordrecht: Kluwer Academic Publisher, 1989. 367 p.

WRIGHT, S. **Evolution and genetics of populations**. 4 vol. Chicago: University of Chicago Press, 1978. 590 p.

WYSMIERSKI, P. T. **Contribuição genética dos ancestrais da soja às cultivares brasileiras**. 2010. 99 f. Dissertação (Mestrado em Ciências: Genética e Melhoramento de Plantas)—Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2010.

## APÊNDICES



**Apêndice A1.** Histograma e quantis normais (QQ-plot) de resíduos da produtividade de grãos ( $\text{kg.ha}^{-1}$ ), resultantes do ajuste de modelo fixo de análise conjunta de variância para experimentos de competição final de genótipos de soja, nos grupos de maturação superprecoce (A), precoce (B) e semiprecoce (C).

**Apêndice A2.** Estimativas REML dos componentes de variâncias residuais ( $\hat{\sigma}_e^2$ ), das médias por localidade ( $PG$ ) e de autocorrelações espaciais de primeira ordem ( $\hat{\rho}$ ), com respectivos p-valores do teste de Durbin-Watson ( $Pr < d$ ) associado à hipótese de ausência de autocorrelação positiva, para análises individuais e conjunta de dados de produtividade de grãos ( $\text{kg.ha}^{-1}$ ), em soja de três grupos de maturação.

| Localidade                           | Superprecoce    |                                 |        |              |          | Precoce |                    |        |              |          | Semiprecoce |                    |        |              |          |
|--------------------------------------|-----------------|---------------------------------|--------|--------------|----------|---------|--------------------|--------|--------------|----------|-------------|--------------------|--------|--------------|----------|
|                                      | GL <sup>1</sup> | $\hat{\sigma}_e^2$ <sup>2</sup> | $PG$   | $\hat{\rho}$ | $Pr < d$ | GL      | $\hat{\sigma}_e^2$ | $PG$   | $\hat{\rho}$ | $Pr < d$ | GL          | $\hat{\sigma}_e^2$ | $PG$   | $\hat{\rho}$ | $Pr < d$ |
| Mineiros-GO                          | 54              | 202828                          | 3987,5 | 0,1815       | 0,0462   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Rio Verde-GO(1)                      | 54              | 240138                          | 3650,8 | 0,3573       | 0,0003   | 78      | 147894             | 3977,0 | 0,0223       | 0,4032   | 83          | 81418              | 3734,0 | 0,1791       | 0,0203   |
| Rio Verde-GO(2)                      | 54              | 78360                           | 3809,8 | 0,0403       | 0,3554   | 77      | 32897              | 1267,5 | 0,0698       | 0,2225   | 84          | 47482              | 1357,1 | 0,0925       | 0,1458   |
| Serranópolis-GO                      | 54              | 122717                          | 3200,9 | 0,0408       | 0,3538   | 78      | 217010             | 3139,5 | 0,1371       | 0,0653   | 78          | 262773             | 2897,8 | 0,1602       | 0,0365   |
| Chapadão do Sul-MS                   | 54              | 182697                          | 3301,1 | -0,1115      | 0,8481   | 60      | 118523             | 3390,9 | 0,0254       | 0,3985   | -           | -                  | -      | -            | -        |
| São Gabriel Oeste-MS                 | 54              | 303634                          | 4078,1 | -0,2220      | 0,9806   | 78      | 146831             | 3894,8 | 0,0828       | 0,1812   | 84          | 163527             | 3631,3 | 0,3119       | 0,0001   |
| Acreúna-GO                           | 54              | 84479                           | 2957,8 | -0,0303      | 0,6096   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Goiatuba-GO                          | 54              | 176891                          | 2850,9 | 0,1786       | 0,0489   | 69      | 476922             | 3286,4 | 0,2179       | 0,0099   | 69          | 384738             | 2946,0 | -0,0752      | 0,7897   |
| Conceição Alagoas-MG                 | 54              | 121304                          | 3337,2 | 0,0298       | 0,3923   | 75      | 122136             | 2968,2 | 0,1413       | 0,0618   | -           | -                  | -      | -            | -        |
| Batatais-SP                          | 54              | 81344                           | 2466,2 | 0,0214       | 0,4221   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Araxá-MG                             | 54              | 260343                          | 3779,2 | 0,0751       | 0,2448   | 77      | 271380             | 3807,4 | 0,0674       | 0,2305   | 84          | 299505             | 4197,9 | 0,2198       | 0,0057   |
| Monte Carmelo-MG                     | 54              | 346575                          | 4123,3 | 0,2190       | 0,0208   | 78      | 312902             | 4667,5 | 0,2049       | 0,0115   | 84          | 418860             | 4835,4 | 0,1261       | 0,0748   |
| Nova Ponte-MG                        | 54              | 144340                          | 3461,9 | 0,0259       | 0,4061   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Cristalina-GO                        | 54              | 86176                           | 3724,0 | 0,0602       | 0,2898   | 76      | 73618              | 3792,4 | 0,0486       | 0,2982   | 84          | 230058             | 3696,6 | 0,1825       | 0,0181   |
| Luziânia-GO                          | 54              | 118886                          | 3234,6 | 0,0693       | 0,2617   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Vianópolis-GO(1)                     | 54              | 132503                          | 2710,8 | 0,1205       | 0,1331   | 78      | 140557             | 2916,1 | 0,0978       | 0,1409   | 84          | 93901              | 2455,1 | 0,0797       | 0,1819   |
| Vianópolis-GO(2)                     | -               | -                               | -      | -            | -        | 77      | 221611             | 3065,1 | 0,2481       | 0,0029   | 84          | 107697             | 2599,2 | 0,1444       | 0,0493   |
| Unaí-MG                              | 54              | 52197                           | 2923,0 | 0,1914       | 0,0378   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Itaberaí-GO                          | 54              | 72767                           | 2361,4 | -0,0542      | 0,6911   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Montividiu-GO                        | 54              | 102602                          | 2961,2 | 0,1302       | 0,1147   | 58      | 124180             | 2938,3 | 0,0741       | 0,0001   | -           | -                  | -      | -            | -        |
| Alto Garças-MT                       | 54              | 225888                          | 2974,6 | -0,1187      | 0,8632   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Campo Verde-MT                       | 54              | 222079                          | 4682,8 | -0,0428      | 0,6529   | 78      | 259395             | 4116,3 | 0,1276       | 0,12755  | 84          | 202878             | 3625,2 | 0,0177       | 0,0177   |
| Itiquira-MT                          | 54              | 97876                           | 4099,8 | 0,1320       | 0,1114   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Primavera Leste-MT                   | 54              | 122755                          | 3607,4 | -0,0597      | 0,7086   | -       | -                  | -      | -            | -        | -           | -                  | -      | -            | -        |
| Análise conjunta                     | 1242            | 155751                          | 3403,7 | 0,0413       | 0,0346   | 1037    | 190924             | 3373,3 | 0,1289       | 0,0001   | 902         | 205288             | 3270,5 | 0,11775      | 0,0001   |
| Teste Bartlett ( $Ba$ ) <sup>3</sup> | -               | 150,25**                        | -      | -            | -        | -       | 184,14**           | -      | -            | -        | -           | 175,15**           | -      | -            | -        |

<sup>1</sup> GL: graus de liberdade associado à estimativa do erro experimental ( $\hat{\sigma}_e^2$ ); <sup>2</sup> Estimativas obtidas pelo modelo BLUE<sub>F</sub> + I; <sup>3</sup>  $Ba$ : estatísticas de Bartlett para homogeneidade de variâncias.