CAPÍTULO 3

MÉTODOS DE ANÁLISE ESTATÍSTICA PARA BLOCOS AUMENTADOS COM APLICAÇÃO NO MELHORAMENTO DE PLANTAS (UMA ABORDAGEM *SAS*® ORIENTADA)

RESUMO

Este trabalho trata da análise estatística para delineamentos em blocos aumentados com ênfase nas suas aplicações ao melhoramento genético vegetal. Procurou-se cobrir diferentes possibilidades de análise, desde o simples ajuste das respostas genotípicas em relação ao desempenho das testemunhas mais próximas e a análise intrablocos, até os procedimentos que permitem recuperar as informações interblocos e intertratamentos. Instruções computacionais em linguagem *SAS* são listadas para a execução das análises, incluindo-se a situação em que os novos genótipos têm origens distintas. A eficiência relativa das diferentes abordagens foi avaliada utilizando-se dados de produtividade de grãos (kg/ha) de 32 ensaios de competição de linhagens de soja (Departamento de Genética, ESALQ/USP). Os resultados indicaram uma maior precisão associada às estimativas dos modelos com aproveitamento de informação inter-efeitos, bem como a possibilidade de seleções diferenciadas entre os procedimentos, sobretudo quando a herdabilidade é baixa. Ademais, as análises que recuperam tal informação, em geral, levam à seleção de uma proporção menor de genótipos em relação à melhor testemunha (mais produtiva) do que os modelos que não a recuperam. São feitos ainda alguns comentários sobre aspectos experimentais desses delineamentos, bem como sobre a natureza dos efeitos de blocos e de genótipos (se fixos ou aleatórios) nesse tipo de experimentação.

Palavras-chave*: delineamentos aumentados, modelos lineares mistos, informação interblocos, informação intergenotípica, informação intervarietal, seleção com base em testemunhas.*

METHODS OF STATISTICAL ANALYSIS FOR AUGMENTED BLOCK DESIGNS WITH APPLICATION TO PLANT BREEDING (A SAS-ORIENTED APPROACH)

ABSTRACT

This work deals with the statistical analysis of augmented block designs emphasizing its applications to plant breeding. Analyses covered different alternatives, ranging from the situations in which selection is based on the behaviour of entries relative to common checks and intrablock analysis, until the analyses with recovery of interblock and intertreatment information. Computational procedures are listed in *SAS* language, including cases in which the new treatments (genotypes) stem from different base populations (e.g. lines from several crosses). The relative efficiency of the different approaches was measures using grain yield data (kg/ha) of soybean inbred lines evaluated in 32 trials (Department of Genetics, ESALQ/USP). In comparison with the intrablock analysis, recovering the interblock and/or intertreatment information lead to: *i)* higher precision of estimates; *ii)* changes in the ranking of lines, meaning that different genotypes would be selected, specially when heritability is low; and, *iii)* smaller fraction of lines ranking above the best check variety. In addition, some comments about experimental aspects of these designs and about the nature of block and genotype effects (fixed *vs*. random) are also made.

Key words*: augmented designs, mixed models, interblock information, intervariety information, intergenotypic information, selection on the checks.*

1. INTRODUÇÃO

A classe dos delineamentos aumentados foi introduzida por Walter T. Federer, na década de cinqüenta. Sua descrição detalhada é apresentada em vários artigos (Federer 1956; 1958; 1961a; 1961b; 1963; Federer & Raghavarao, 1975; Federer *et al.*, 1975). A proposta veio atender a uma necessidade experimental presente nas fases preliminares dos programas de melhoramento vegetal, quando o número de entradas a serem avaliadas é bastante alto e a quantidade de material de propagação (sementes, tubérculos, etc.) para cada uma delas é pequena. Um delineamento aumentado é obtido escolhendo-se um delineamento experimental padrão para os tratamentos controle (ou testemunhas), aumentando-se, em seguida, os seus blocos (linhas e/ou colunas) em parcelas que acomodarão os tratamentos adicionais. Estes últimos tratamentos (novos genótipos), usualmente aparecem uma só vez em todo o experimento, embora tal número não seja uma exigência desses delineamentos. As testemunhas podem estar dispostas em blocos completos ou incompletos, balanceados ou parcialmente balanceados. Mas, o conjunto completo dos tratamentos (testemunhas + tratamentos novos) sempre estará distribuído em blocos incompletos, ou seja, segundo um delineamento não ortogonal. O mesmo ocorre quando se consideram apenas os novos tratamentos (Federer, 1998). Outra característica típica desses ensaios, geralmente de grande extensão, é a presença de algum desbalanceamento não planejado; seja em decorrência da perda de parcelas, seja pela repetição desigual de alguns tratamentos genéticos (ex: progênies com maior quantidade de sementes).

A proposta inicial de Federer (1956) buscava principalmente a obtenção de médias de tratamentos ajustadas para os efeitos de blocos, uma vez que os novos tratamentos, em sua maior parte, são testados entre si em blocos diferentes. Assim, o enfoque foi dirigido basicamente para o modelo de efeitos fixos (blocos e tratamentos), ou seja, restringindo-se à denominada *análise intrablocos*. Este tipo de análise pode ser tratada segundo a abordagem geral para delineamentos em blocos, já amplamente descrita (John, 1971; 1980; Nigam *et al.*, 1989; Iemma, 1987; entre outros). Por outro lado, nesses delineamentos, a ausência de repetições para a maior parte dos tratamentos faz por merecer a introdução de todo tipo de procedimento que possa trazer algum benefício à análise. Um deles é a possibilidade de admitir os efeitos de blocos como aleatórios, com conseqüente melhoria da eficiência da análise através da chamada *recuperação da informação interblocos* (Yates, 1939; Rao, 1947; Patterson & Thompson, 1971). Usualmente, este tipo de análise (com blocos aleatórios) também é aplicado admitindo-se os efeitos de tratamentos como fixos. Todavia, como afirmam Wolfinger *et al.* (1997) e Federer (1998), nos estágios iniciais dos

programas de seleção, a natureza aleatória dos genótipos sob teste também precisa ser levada em consideração. Isso permite, adicionalmente, recuperar o que estes autores denominam de *informação intervarietal* ou *intergenotípica*, seja isoladamente ou em combinação com a informação interblocos. Dessa forma, quatro opções de análise tornam-se disponíveis, a primeira baseada em modelo fixo e as demais na abordagem de modelos mistos.

No Programa de Melhoramento da Soja desenvolvido no Departamento de Genética da ESALQ/USP, o delineamento de blocos aumentados vem sendo utilizado desde 1991. O processo rotineiro de seleção, entretanto, tem sido feito com base no desempenho de cada progênie em relação à *performance* das testemunhas próximas, pertencentes ao mesmo bloco ou conjunto. Este método, doravante chamado *seleção com base em testemunhas intercalares*, sabidamente, não fornece uma estimativa do erro experimental (Wolfinger *et al*., 1997). Ademais, não permite inferências estatísticas acerca da(s) população(ões) de que os genótipos se originaram (ex: estimação da variância genética, herdabilidade, predição de valores genotípicos, etc.). Entre as razões para a adoção desse procedimento está a suposta dificuldade de implementar uma análise estatística mais eficiente, com a rapidez operacional que o processo exige. Por isso, no contexto do referido programa, análises mais apuradas têm-se restringido a trabalhos de tese (Farias Neto, 1995; Gomes, 1995; Laínez-Mejía, 1996; Azevedo Filho, 1997; Pinheiro, 1998; Hamawaki, 1998). Ainda assim, o tratamento estatístico adotado, no que se refere à determinação das médias de tratamentos, tem-se limitado à análise intrablocos. Dessa forma, o estabelecimento de rotinas computacionais que possibilitem a aplicação rotineira de critérios seletivos de maior rigor estatístico é uma necessidade do Programa; haja vista a debilidade dos delineamentos aumentados em termos de precisão, sobretudo para contrastes de grande interesse como os que envolvem os novos genótipos.

Até o início da década de oitenta, as limitações de recursos computacionais contribuíram de forma decisiva para a adoção dos modelos fixos. Isso porque a abordagem estatística através dos chamados modelos lineares mistos exige um esforço computacional consideravelmente maior (Luengo & Martín, 1995; Bueno Filho, 1997). Atualmente, com a ampla disponibilidade de *softwares* estatísticos (*SAS*, *STATISTICA*, *MINITAB*, *GENSTAT*, etc.), a adoção da análise intrablocos, por motivos dessa natureza, não mais se justifica. Assim, este tipo de análise deve-se restringir às situações em que a suposição de aleatoriedade de certos efeitos (blocos e/ou tratamentos adicionais) não puder realmente ser assumida. E mesmo nestas situações, há quem considere o uso da modelagem mista sempre conveniente para a melhoria da eficiência das análises (Federer, 1998; Piepho, 1994); haja vista a possibilidade de levar em conta alguma dependência

(covariância) entre os níveis de fatores aleatórios que não seria considerada num modelo fixo, o que geralmente implica num aumento da precisão das estimativas.

Diante desse quadro, o propósito deste trabalho foi compilar os avanços mais recentes na análise de delineamentos aumentados, incluindo recuperação de informações interblocos e intervarietal, procurando desmistificar a sua aplicação. Mais especificamente, além de alguma fundamentação teórica sobre as opções (modelos) de análise, o trabalho procura oferecer aos usuários procedimentos computacionais que possibilitem a sua pronta execução. As instruções são apresentadas para uso através do sistema estatístico *SAS®* (*Statistical Analysis System*). Tal escolha baseou-se na sua popularidade e no fato de permitir uma programação através de poucas linhas de comandos, graças a seus procedimentos internos (*PROC'*s) previamente programados. O trabalho procura também comparar as alternativas de análise, buscando identificar as situações em que as abordagens diferentes produzem resultados mais discrepantes. Com isso pode-se alertar os usuários para os riscos de uma especificação não apropriada do modelo de análise. Ademais, enumera-se algumas recomendações sobre o planejamento desse tipo de ensaio, com vistas à obtenção de uma análise estatística de maior eficiência.

2. METODOLOGIA

2.1. Material

O material do presente estudo refere-se a conjuntos de dados de 32 experimentos delineados em blocos aumentados. Os ensaios, provenientes do Programa de Melhoramento da Soja desenvolvido pelo Setor de Genética Aplicada às Espécies Autógamas, do Departamento de Genética da ESALQ/USP, foram conduzidos entre os anos de 1992 e 1996 (Tabela 3.1, na seção Anexos). Mais especificamente, os experimentos fazem parte do programa de seleção recorrente visando o aumento da produtividade de grãos na soja. Os tratamentos compreendem, portanto, as testemunhas, variando entre 4, 5 e 10, e um número variável de progênies (entre 60 e 1260), correspondentes aos tratamentos adicionais.

As progênies foram resultantes de cruzamentos biparentais (subprogramas Precoce, Semi-precoce e Semi-tardio) ou de intercruzamentos envolvendo quatro pais (subprogramas Precoce *x* Semi-precoce e Semi-tardio *x* Tardio), em conformidade com a filosofia de seleção recorrente para espécies autógamas, adotada no Programa (Vello, 1992; Frey, 1976). Todo o material genético, por ocasião dos ensaios, já se encontrava em gerações avançadas de endogamia, variando entre $F_{6:2}$ a $F_{11:6}$ (Tabela 3.1). Informações detalhadas a respeito dos parentais envolvidos, composição de

germoplasma exótico no conjunto desses genitores, esquemas de cruzamentos e métodos de avanço de gerações podem também ser buscadas em Alliprandini (1996) e Lopes (1996).

Embora diversos caracteres de interesse para o melhoramento da soja tenham sido avaliados, apenas os dados de produtividade de grãos (kg/ha) foram aqui considerados. Ademais, dada a perda de parcelas e a necessidade de descarte de observações discrepantes (*outliers*), com vistas à garantia de distribuição normal dos dados, os conjuntos analisados nem sempre apresentaram exatamente os tamanhos descritos na Tabela 3.1. Deve-se informar que alguns ensaios foram descartados no início do estudo, em função de coeficientes de variação elevados (CV>35%) na análise intrablocos, restando, então, os 32 anteriormente referidos.

É importante ressaltar que os dados experimentais foram utilizados, no presente trabalho, apenas como material para avaliação das diferentes abordagens analíticas, e não com o propósito de, efetivamente, selecionar genótipos para o desenvolvimento do Programa. Primeiramente, porque isto já fora feito no devido tempo, haja vista que se trata de ensaios conduzidos há mais de três anos. Em segundo lugar, porque, na maioria deles, as linhagens são oriundas de vários cruzamentos (procedências) e este fato não foi levado em conta nos modelos de análise aqui avaliados.

2.2. Modelos alternativos de análise estatística

Os métodos estatísticos implementados compreendem, na realidade, procedimentos alternativos para a seleção de progênies, em ensaios de competição delineados em blocos aumentados. Entre os procedimentos utilizados, enfatizou-se os modelos de análise com maior rigor estatístico: *análise intrablocos*; *análise com recuperação da informação interblocos*; *análise com recuperação de informação intervarietal*; e *análise recuperando ambos os tipos de informação*, *interblocos e intervarietal*. Todas estas opções podem ser descritas, inicialmente, a partir do modelo geral de delineamentos em blocos, tal como em Boyle & Montgomery (1996) e Marcos (1994):

$$Y_{ij} = \mu + \beta_j + \tau_i + \varepsilon_{ij} \qquad \text{(I)}$$

em que:

Y_{ij}: é a resposta observada do *i-ésimo* tratamento no *j-ésimo* bloco (totalizando *n* observações);

μ : é a constante comum às observações (uma referência à média geral das observações);

β_j : é o efeito do *j-ésimo* bloco (*j=1,2,...,b*);

τ_i : é o efeito do *i-ésimo* tratamento (*i=1, 2, ..., p, p+1, p+2, ..., p+t*; sendo *p* o número de progênies ou novos tratamentos, *t* o número de testemunhas e *p+t=v*, o número total de tratamentos); e

ε_{ij} : é o erro experimental aleatório associado à parcela com o *i-ésimo* tratamento, no *j-ésimo* bloco, distribuído *normal* e independentemente, com média zero e variância σ_e^2 ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$).

A partir deste modelo geral, as análises implementadas (modelagens alternativas) podem ser diferenciadas conforme as suposições a seguir:

i) *Modelo 1 (Fixo)*: efeitos fixos para blocos e tratamentos, correspondendo à análise intrablocos. Os efeitos μ, β_j, τ_i e ε_{ij} são admitidos independentes entre si e o único componente de variância está associado ao erro experimental, conforme a suposição: $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

ii) *Modelo 2 (Misto A)*: efeitos aleatórios para blocos e fixos para tratamentos, correspondendo à chamada análise com recuperação da informação interblocos. Além da suposição de independência entre μ, β_j, τ_i e ε_{ij}, admite-se: $\beta_j \sim N(0, \sigma_b^2)$ e $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

iii) *Modelo 3 (Misto B)*: efeitos fixos para blocos e aleatórios para tratamentos, exceto testemunhas (de efeitos sempre fixos), o que corresponde a uma análise com recuperação de informação intervarietal (intergenotípica). Além da independência entre μ, β_j, τ_i e ε_{ij}, admite-se: $\tau_i \sim N(0, \sigma_g^2)$, com $i=1,2,...,p$ e $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

iv) *Modelo 4 (Misto C)*: efeitos aleatórios para blocos e tratamentos (exceto testemunhas), correspondendo a uma análise com recuperação das informações interblocos e intervarietal. Assume-se também independência entre μ, β_j, τ_i e ε_{ij}, e efeitos aleatórios distribuídos conforme: $\beta_j \sim N(0, \sigma_b^2)$; $\tau_i \sim N(0, \sigma_g^2)$, com $i=1,2,...,p$; e $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

Para avaliar as conseqüências dos diferentes modelos e suas respectivas suposições sobre o processo de seleção, todos os conjuntos de dados foram submetidos aos quatro tipos de análise. É necessário ressaltar que, na prática (realidade), cada conjunto de dados (experimento) admite apenas um destes modelos analíticos. Assim, o exercício estatístico aqui desenvolvido é feito no sentido de apontar as possíveis disparidades produzidas pelas modelagens alternativas. Com isso, pode-se alertar os usuários sobre os riscos de uma especificação inadequada do modelo de análise, haja vista a possibilidade de resultados (testes estatísticos, estimativas de médias de tratamentos e respectivo ordenamento) bastante discrepantes, com conseqüente tomada de decisões equivocadas.

Dada a existência de duas categorias ou tipos de tratamentos (*testemunhas* e *tratamentos novos*), o modelo geral de delineamentos em blocos não possibilita a análise através dos *Modelos 3* e *4*, em que os efeitos de tratamentos são de duas naturezas: fixos para as testemunhas e aleatórios para as progênies. Uma alternativa é reparametrizar os efeitos τ_i numa estrutura hierárquica, de forma que os novos tratamentos representem efeitos aleatórios distribuídos em torno de uma média fixa comum, relativa à população da qual foram retirados; enquanto as testemunhas, de efeitos fixos, representem outras *t* médias paramétricas diferentes (Scott & Milliken, 1993; Wolfinger *et al*., 1997; Federer, 1998). O modelo proposto por Scott & Milliken (1993) é:

$$Y_{ij} = \mu + \beta_j + c_i + x_i(c_i) + \varepsilon_{ij} \qquad \text{(II)}$$

Em que, além dos termos já definidos, c_i e $x_i(c_i)$ denotam o efeito de tratamento, com $c_i + x_i(c_i) = \tau_i$, em relação ao modelo (I), e, se *i* for uma testemunha tem-se: $x_i(c_i) = 0 \Rightarrow c_i = \tau_i$. A variável de classificação ou fator *C*, associada aos efeitos c_i, é definida como a identificação do tratamento *i* (nome do genótipo), se este for uma testemunha, e *'0'* (zero) se *i* for uma progênie (novo tratamento). Já o fator *X*, associado aos efeitos x_i, recebe a identificação do tratamento se *i* for uma progênie e *'0'* se *i* for uma testemunha.

A nova expressão possibilita, pelo menos em princípio, incorporar os quatro modelos de análise apresentados. O fator *C* possui, então, um nível a mais (o nível *zero*) do que o número de testemunhas, perfazendo *t+1* populações. Neste caso, os efeitos c_i são considerados sempre fixos, pois referem-se a desvios médios populacionais (c_0, c_1, c_2, ..., c_t). O efeito c_0 corresponde à população que deu origem às progênies, a qual é também de natureza fixa em relação às testemunhas. A fonte de variação associada a este fator terá, então, tantos graus de liberdade quantas forem as testemunhas, pois incorpora as variações: "entre testemunhas", com *(t−1) GL*, mais "testemunhas *vs*. progênies", com *1 GL*. Assim, mesmo que a variação total em *C* não desperte interesse prático por si só, ela pode ser desdobrada nestes contrastes ortogonais de alguma importância para o melhorista.

O efeito de progênie dentro da população, $x_i(c_i)$, é que pode, então, ser de natureza fixa (*Modelos 1* e *2*) ou aleatória (*Modelos 3* e *4*). Como há apenas um nível dentro das populações de testemunhas (*0 GL*), toda a variação em *X(C)* é devida à variabilidade "entre progênies"; ou seja, dentro do nível '*0*' de *C* (com *p−1 GL*). Portanto, esta fonte de variação reflete, de fato, somente os efeitos de progênies dentro da população que lhes deram origem. Assim, ao assumir os efeitos $x_i(c_i)$ como aleatórios obtém-se as análises de variância apropriadas para os *Modelos 3* e *4*. Por outro lado, nota-se que o fator *X* possui *t* níveis a mais do que o número de progênies (referentes às testemunhas), o que acrescenta dificuldades computacionais à estimação dos erros padrão de médias. E, atualmente, os sistemas de análise estatística, a exemplo do *SAS*, não permitem particionar uma variável, como tratamentos, num subconjunto que é fixo (as testemunhas) e noutro que é aleatório (as progênies). Assim, na prática, ainda é necessário construir variáveis auxiliares que possibilitem incorporar este fato (Wolfinger *et al.*, 1997), o que será apresentado no item *2.3*.

Os parâmetros de maior interesse neste estudo compreenderam os componentes de variância relacionados a cada modelo de análise, as médias de tratamentos (testemunhas e progênies) e seus erros padrão associados ($s_{\bar{Y}}$). Nos delineamentos aumentados, em função de desbalanceamento(s), é natural obter-se erros de média diferentes de um tratamento para outro.

Porém, considerando-se o interesse voltado à comparação dos modelos de análise deu-se preferência, aqui, aos erros padrão médios, tomados, separadamente, para testemunhas ($\overline{s}_{\overline{Y}\mathrm{T}}$) e progênies ($\overline{s}_{\overline{Y}\mathrm{P}}$). Com as estimativas dos parâmetros foram ainda obtidas algumas relações importantes para a comparação dos modelos: *i)* $\phi_b = \hat{\sigma}_b^2 / \hat{\sigma}_e^2$ ou/e $\phi_g = \hat{\sigma}_g^2 / \hat{\sigma}_e^2$ (razões de variâncias, nos *Modelos 2* ou *3* e *4*, respectivamente); *ii)* $(\overline{s}_{\overline{Y}\mathrm{P}} / \overline{s}_{\overline{Y}\mathrm{T}})100$, o erro padrão médio (%) de progênies em relação ao de testemunhas para cada um dos modelos; e *iii)* $(\overline{s}_{\overline{Y}\mathrm{P}} / \overline{s}_{\overline{Y}\mathrm{P}}^{\mathrm{Fixo}})100$, o erro padrão médio (%) de progênies de cada modelo em relação ao do modelo fixo.

Na comparação das médias ajustadas pelas diferentes análises, bem como dos respectivos ordenamentos produzidos, dois outros procedimentos seletivos foram incluídos. O primeiro, referido como *Modelo 0*, consistiu na obtenção das médias marginais dos tratamentos (médias simples, não ajustadas). O segundo baseou-se na seleção com base em testemunhas usada rotineiramente na condução do Programa. Neste procedimento, referido como *Modelo 5*, converteu-se a produtividade de cada progênie (Y_{ij}) numa porcentagem em relação à média das *testemunhas*, no bloco em que a progênie apareceu. Se a progênie ocorreu em mais de um bloco, a produtividade percentual é dada pela média das porcentagens intrablocos. Estes dois procedimentos, entretanto, foram utilizados sobretudo para estabelecer o ordenamento das progênies para fins de seleção.

A partir disso, foram estimados ainda os coeficientes de correlação linear de Pearson (entre as médias de tratamentos obtidas pelos modelos alternativos) e de Spearman (entre os respectivos postos destas médias). Avaliou-se também a dispersão das estimativas das médias de progênies ($\overline{Y}_i$), obtidas em cada procedimento, por: $\mathrm{CV_P} = \frac{100}{\hat{\mu}_\mathrm{P}} \sqrt{\sum_{i=1}^{p} (\overline{Y}_i - \hat{\mu}_\mathrm{P})^2 / (p-1)}$; sendo $\hat{\mu}_\mathrm{P}$ a estimativa da média geral das progênies. Adicionalmente, para avaliar os efeitos das diferentes abordagens sobre a seleção com base em testemunhas, determinou-se o número de progênies e a porcentagem dos genótipos cujas médias superaram a da testemunha superior ($\hat{\mu}_{(\mathrm{T.sup})}$). Por fim, foi escolhido um dos experimentos para ilustrar numericamente algumas conseqüências dos métodos estatísticos utilizados.

2.3. Procedimentos estatístico-computacionais

As análises foram implementadas através do sistema estatístico-computacional *SAS*® utilizando-se, em especial, seus procedimentos para análise de modelos lineares: o *PROC GLM* (*procedure for general linear models*) e o *PROC MIXED* (*procedure for mixed linear models*). As rotinas computacionais foram construídas adaptando-se uma série de instruções já disponíveis para

esses tipos de análises (Scott & Milliken, 1993; Marcos, 1994; Boyle & Montgomery, 1996; Federer & Wolfinger, 1996; Wolfinger *et al.*, 1997; Piepho, 1997; Federer & Wolfinger, 1998; SAS Institute, 1996; Littell *et al*., 1996; SAS Institute, 1997).

2.3.1. Preparação e leitura do arquivo de dados

O passo inicial para a execução de uma análise estatística é, sem dúvida, a preparação do arquivo de dados. Embora o sistema *SAS* permita a entrada de dados simultaneamente às instruções da análise, prefere-se, aqui, orientar o usuário para a montagem do arquivo numa planilha eletrônica (*EXCEL*, *QUATTRO PRO*, etc.); seja em razão do grande tamanho destes arquivos, seja pelas facilidades oferecidas por estes aplicativos. O arquivo deve conter as observações de cada parcela experimental em suas linhas, identificadas por colunas indicadoras dos respectivos: bloco, tratamento (nome do genótipo), tipo de tratamento (T: testemunhas e P: progênie) e variável(veis) resposta(s). A seguir é apresentado um exemplo de como pode ser montado o arquivo, com: $b=3$ blocos, de tamanhos diferentes ($k_1=7$, $k_2=6$ e $k_3=5$); $t=2$ testemunhas (**BOSSIER** e **IAC-12**); $p=9$ progênies (linhagens experimentais **USP**'s); e, duas variáveis respostas (**APM**: altura da planta à maturidade, em centímetros, e **PG**: produtividade de grãos, em kg/ha):

BLOCO	GENÓTIPO	TIPO	APM	PG
1	IAC-12	T	101	1275,0
1	USP 93-11040	P	85	1540,4
1	USP 94-10025	P	100	1786,1
1	USP 93-08047	P	89	984,7
1	BOSSIER	T	95	1824,5
1	USP 93-09083	P	87	1036,7
1	IAC-12	T	103	1435,0
2	USP 93-11510	P	100	2153,8
2	IAC-12	T	78	1976,3
2	USP 94-11023	P	94	897,5
2	BOSSIER	T	97	1976,4
2	USP 94-10025	P	99	1247,9
2	USP 93-09815	P	85	721,6
3	USP 94-11023	P	100	1203,8
3	USP 93-09911	P	83	1158,8
3	IAC-12	T	77	1092,7
3	BOSSIER	T	96	1733,9
3	USP 94-11201	P	102	1281,6

O arquivo assim montado numa planilha *EXCEL*, por exemplo, deve, então, ser gravado em formato texto, separado por espaços (ASC II: com extensão '**.PRN**'). Em seguida, este deve ser aberto no sistema *SAS*, quando deverá ser eliminada a sua primeira linha (**BLOCO GENÓTIPO** ... **PG**) e, se necessário (no caso da planilha em língua portuguesa), trocar '**,**' (vírgula) por '**.**' (ponto), como separador de decimal. Para isso, deve-se marcar todo o conjunto de dados (no *SAS*) e utilizar a seqüência de comandos: Edit→Replace→(Find What: **,** → Replace With: **.**)→Replace All. Após tais

ajustes de compatibilidade, grava-se novamente o arquivo (ex: **ARQUIVO.PRN**), o qual estará pronto para ser utilizado pelo *SAS*.

O passo seguinte consiste na leitura do arquivo de dados pelo sistema *SAS*. Para isso é necessária uma seqüência de comandos como a apresentada adiante. Aqui, termos com letras minúsculas indicam comandos e variáveis internas do *SAS* e os com letras maiúsculas são nomes fornecidos pelo usuário, embora o sistema não faça qualquer exigência nesse sentido:

```
data ARQ_ORIG;
infile 'C:\DADOS\ARQUIVO.PRN';
input BLOCO GENOT$ TIPO$ APM PG;
  C=GENOT; if TIPO^='T' then C='0';
  X=GENOT; if TIPO='T' then X='0';
  if TIPO='P' then NEW=1; else NEW=0;
run;
```

Executando-se este programa o sistema reconhece o arquivo como um *dataset SAS*, aqui denominado **ARQ_ORIG**, que passa a fazer parte de seu diretório de trabalho (*SASWORK*). Em seguida, pode-se submeter o conjunto de dados aos seus diversos procedimentos de análise (*PROC GLM*, *PROC MIXED*, etc.).

2.3.2. Instruções para a execução das análises estatísticas

Análise intrablocos (Modelo 1 - Fixo)

A análise de blocos aumentados com base num modelo fixo (análise intrablocos) pode ser implementada, no *SAS*, através de uma variedade de instruções. Marcos (1994) e Boyle & Montgomery (1996) apresentam seqüências de comandos baseadas no modelo (I) e Scott & Milliken (1993) listam um programa com base no modelo (II). Ambas as abordagens não fornecem o desdobramento natural para a fonte de variação 'Tratamentos', ou seja, as variações devidas a 'Testemunhas', 'Progênies' e 'Testemunhas *vs*. Progênies'; a menos que sejam construídas, manualmente, as matrizes dos contrastes correspondentes para uso junto ao comando '**contrast**' dos *PROC*'s *GLM* ou *MIXED*. Especialmente no caso do modelo geral de delineamentos em blocos (I), isso pode representar uma tarefa laboriosa quando o número de tratamentos for elevado (maioria desses ensaios). Nesse sentido, a proposta de Scott & Milliken (1993) trouxe uma grande contribuição, pois permite reduzir sensivelmente a dimensão das matrizes de contrastes. As seqüências de comandos sugeridas pelos dois grupos de autores são:

Marcos (1994) e Boyle & Montgomery (1996):

```
proc glm data=ARQ_ORIG;
  class BLOCO GENOT;
  model PG=BLOCO GENOT;
  lsmeans GENOT/stderr;
run;
```

Scott & Milliken (1993):

```
proc glm data=ARQ_ORIG;
  class BLOCO C X;
  model PG=BLOCO C X(C);
  lsmeans X(C)/stderr;
run;
```

Ambos os programas produzem análises equivalentes (mesmas médias ajustadas, erros padrão, etc.), embora os efeitos de tratamentos estejam diferentemente representados (**GENOT** e **C**+**X(C)**, respectivamente). Assim, as duas análises podem ser combinadas para a composição de um quadro de *ANOVA* mais completo e informativo:

F.V.	GL
BLOCO	$b-1$
GENOT	$v-1=t+p-1$
C	t
X(C)	$p-1$
Erro	$n-(b+v-2)$
Total	$n-1$

Vale relembrar que a fonte de variação '**C**' incorpora os efeitos: 'Testemunhas', com *(t-1) GL*, e 'Testemunhas *vs*. Progênies', com *1 GL*. Para obter este desdobramento é necessário, então, construir as matrizes associadas ao comando '**contrast**'. Na proposta de Scott & Milliken (1993), isso pode ser feito adicionando-se, após a linha '**model** ... **;**', as instruções a seguir (o primeiro nível de '**C**' corresponde às progênies e admite-se, como exemplo, *t=3* testemunhas):

```
contrast 'TESTEMUNHAS'  C  0  1 -1  0,
                        C  0  0  1 -1;
contrast 'TEST vs PROG' C -3  1  1  1;
```

Para evitar o trabalho de construção das matrizes e possíveis equívocos aos menos familiarizados com as peculiaridades do comando '**contrast**', apresenta-se aqui uma modelagem alternativa também equivalente. Dada a classificação dos genótipos em duas categorias, testemunhas e progênies, o modelo de análise ainda pode ser escrito como:

$$Y_{ijk} = \mu + \beta_j + T_k + g_{i(k)} + \varepsilon_{ij} \qquad \text{(III)}$$

em que, além de efeitos já definidos:

T_k : é o efeito do *k-ésimo Tipo* de tratamento, com *k=1* ou *2*, conforme o tratamento seja uma testemunha (**T**) ou uma progênie (**P**), respectivamente; e

$g_{i(k)}$: é o efeito do genótipo *i* dentro do *k-ésimo* tipo de tratamento.

Considerando-se que o arquivo de dados já foi montado com a variável classificatória '**TIPO**', esta análise pode ser obtida com a seguinte instrução:

```
proc glm data=ARQ_ORIG;
  class BLOCO TIPO GENOT;
  model PG=BLOCO TIPO GENOT(TIPO);
  lsmeans GENOT(TIPO)/slice=TIPO stderr cl out=MEDIAS;
run;
```

Nesta alternativa, a fonte de variação '**TIPO**' corresponde diretamente ao contraste 'Testemunhas *vs* Progênies'. Já '**GENOT(TIPO)**', assim como '**C**' na proposta de Scott & Milliken (1993), não tem interesse prático direto (variação genotípica média dentro das duas populações ou tipos). Contudo, o desdobramento de '**GENOT(TIPO)**', em '**T**' ($k=1$) e '**P**' ($k=2$), propiciado pela opção '**slice**', corresponde, respectivamente, aos contrastes entre 'Testemunhas' e entre 'Progênies', ambos de grande interesse. Este último é equivalente à fonte de variação '**X(C)**' da proposta anterior (variação entre genótipos dentro da população de progênies). Dessa forma, os principais testes estatísticos podem ser obtidos sem a necessidade da construção manual de matrizes, evitando-se, adicionalmente, possíveis equívocos de interpretação relacionados à ordem dos níveis considerada na função '**contrast**'. Logo, este programa propicia a montagem direta do quadro de *ANOVA*, em conformidade com os interesses da maioria dos melhoristas, ou seja:

F.V.	GL
BLOCO	$b-1$
TIPO (Test. *vs* Prog.)	1
GENOT(TIPO)	$v-2=t+p-2$
T (Testemunhas)	$t-1$
P (Progênies)	$p-1$
Erro	$n-(b+v-2)$
Total	$n-1$

Vale ressaltar que a execução da análise pelo modelo (III), ainda não descrita na literatura de blocos aumentados, só ganhou aplicabilidade graças à opção '**slice**' do comando '**lsmeans**', disponível a partir da versão 6.11 do *SAS*. Tendo em vista a disponibilidade atual de versões *SAS*® igual ou superior à 6.11, esta última modelagem analítica apresenta, portanto, vantagens práticas comparativas. Um fato que provavelmente contribuiu para retardar a aplicação desta modelagem através do sistema *SAS*, é que a opção '**slice**' foi apresentada para aplicação em experimentos fatoriais, com o seguinte formato: '**lsmeans A*B/slice=B;**' (SAS Institute, 1997). Esta sintaxe sugere que a opção '**slice**' (cortar em fatias) processa um desdobramento da interação '**A*B**', o que não é completamente correto. Pois, na realidade, o que é feito através desta instrução é

o desdobramento de **A(B)**; ou seja, o total de graus de liberdade desdobrados pela opção corresponde a: *GL*(**A**)+*GL*(**A*B**) e não somente a *GL*(**A*B**), o mesmo verificando-se para o desdobramento de somas de quadrados. Isto adequa-se perfeitamente aos objetivos da análise em questão, tornando a parte programável de sua execução bastante simples e passível de uso rotineiro nos programas de seleção.

Além das especificações já descritas para o programa *SAS* anterior, o comando '**lsmeans**' invoca a obtenção das estimativas das médias de tratamentos, ajustadas por quadrados mínimos, bem como o erro padrão associado a cada uma (opção '**stderr**'). Também permite o cálculo do respectivo intervalo de confiança, via opção '**cl**' (o nível de confiança padrão é 95%, mas pode ser alterado pela opção '**alpha=**'). Por fim, a opção '**out=MEDIAS**' produz um arquivo *dataset SAS* que armazena todas essas estimativas.

É oportuno mencionar que essas rotinas de análise podem ser implementadas, com algumas pequenas alterações, utilizando-se o *PROC MIXED*, ao invés do *PROC GLM*. Isso pode ser consideravelmente vantajoso em termos de redução no tempo de processamento da análise, no caso de experimentos muito grandes (acima de 500 progênies). Embora o formato de apresentação dos resultados pelo *PROC MIXED* não seja o mais costumeiro, com quadro de *ANOVA*, etc., todas as saídas do *PROC GLM* estão de alguma forma disponíveis. Então, para a análise de modelo fixo, a seqüência de comandos é praticamente a mesma utilizada no *PROC GLM*, a menos da substituição óbvia de '**proc glm**' por '**proc mixed**' e a exclusão dos termos '**stderr**' e '**out=MEDIAS**' como opções do comando '**lsmeans**'. A dispensa do primeiro termo justifica-se porque, no *PROC MIXED*, os erros padrão são listados automaticamente a partir de '**lsmeans**'. Já o *dataset SAS* que armazena as médias ajustadas e demais estatísticas relacionadas, é gerado a partir de outra linha de comandos: '**make 'lsmeans' out=MEDIAS;**', a ser inserida após a linha '**lsmeans** ...'.

Outro procedimento computacional que geralmente interessa aos melhoristas é a ordenação decrescente das médias de tratamentos para fins de seleção. Isso pode ser feito facilmente executando-se as seguintes instruções (*dataset* '**MEDIAS**' obtido pelo *PROC GLM*):

```
proc sort data=MEDIAS;
  by descending lsmean;
proc print; run;
```

Obs.: Nesta rotina, se o *dataset* '**MEDIAS**' for obtido pelo *PROC MIXED*, em versões *SAS* iguais ou superiores à 6.12, a variável interna '**lsmean**' deve ser informada como '**_lsmean_**'.

Análise com recuperação de informação interblocos (Modelo 2 – Misto A)

Na análise com recuperação da informação interblocos, os efeitos de blocos são considerados aleatórios, caracterizando-se um modelo misto com dois componentes de variância, σ_b^2 e σ_e^2. Como no caso de modelo fixo, aqui também pode-se adotar quaisquer das opções de modelagem apresentadas anteriormente (I, II ou III), com a diferença na suposição dos efeitos de blocos, agora aleatórios. Assim, as instruções para execução da análise segundo as duas primeiras propostas podem ser listadas como:

Marcos (1994) e Boyle & Montgomery (1996):

```
proc mixed data=ARQ_ORIG;
  class BLOCO GENOT;
  model PG=GENOT;
  random BLOCO;
  lsmeans GENOT/cl;
run;
```

Scott & Milliken (1993):

```
proc mixed data=ARQ_ORIG;
  class BLOCO C X;
  model PG=C X(C);
  random BLOCO;
  lsmeans X(C)/cl;
run;
```

Como no caso de modelo fixo, também aqui a modelagem III pode ser vantajosa, pois permite obter mais prontamente os testes para contrastes de interesse:

```
proc mixed data=ARQ_ORIG method=reml;
  class BLOCO TIPO GENOT;
  model PG=TIPO GENOT(TIPO);
  random BLOCO;
  lsmeans GENOT(TIPO)/slice=TIPO cl;
  make 'lsmeans' out=MEDIAS;
run;
```

A substituição do *PROC GLM* pelo *PROC MIXED*, nas três opções, é uma necessidade da abordagem de modelo linear misto, própria da análise em questão. O *PROC GLM*, entretanto, continua podendo ser usado especialmente para a montagem de quadros de *ANOVA* e para a obtenção das esperanças de quadrados médios (via comando '**random**'). Todavia, resultados de médias ajustadas e erros padrão associados, bem como alguns testes estatísticos, não são listados corretamente. Como informa o SAS Institute (1993b), as propriedades inerentes ao *PROC GLM* assumem um modelo fixo, assim, os testes estatísticos padrão, incluindo as verificações sobre estimabilidade, não são apropriadas para modelos mistos, mesmo quando se usa o comando '**random**'. Portanto, da saída padrão (*default*) do *PROC GLM* somente devem ser aproveitados os resultados relativos a: *GL*, *SQ* e *QM*. Os testes corretos são obtidos somente quando se utiliza, adicionalmente, a opção '**test**' do comando '**random**'; recomendando-se, neste caso, uma inspeção cuidadosa das *E(QM)*'s e dos testes de hipóteses, por fonte de variação, incluindo-se a respectiva indicação do termo de *erro* apropriado (*GL* e *QM* do denominador da estatística *F*).

Por outro lado, a sintaxe para o *PROC MIXED* é praticamente idêntica à utilizada para o *PROC GLM*, não acarretando, portanto, dificuldades adicionais. Algumas exceções importantes são: *i)* apenas os efeitos fixos devem ser listados no comando '**`model`**' do *PROC MIXED*, enquanto no *PROC GLM* todos os efeitos devem ali aparecer (mesmo com o uso simultâneo do comando '**`random`**'); *ii)* o comando '**`lsmeans`**' lista automaticamente os erros padrão, dispensando-se a opção '**`stderr`**'; e *iii*) o *PROC MIXED* trabalha diretamente os efeitos aleatórios, estimando seus componentes de variância, os quais são levados em conta no ajuste interblocos das médias de tratamentos, na obtenção dos erros padrão e de estatísticas *t* associadas (Wolfinger *et al*., 1997). Na estimação dos componentes de variância o *PROC MIXED* utiliza como *default*, o método de máxima verossimilhança restrita (*REML*). Mas, dois outros métodos também estão disponíveis, o de máxima verossimilhança (*ML*) e o de estimação quadrática não viesada de variância mínima (*MIVQUE-0*). A escolha por um deles é feita entrando-se com a sigla correspondente na opção '**`method=`**', conforme ilustrado na rotina anterior através da escolha '**`reml`**' (especificação supérflua, neste caso, pois *REML* já é o método *default* do procedimento). A adoção do método *REML*, no presente trabalho, decorre de recomendações favoráveis à sua utilização no caso de modelos mistos com dados desbalanceados (SAS Institute, 1996; Verneque, 1994).

Análise com recuperação de informação intergenotípica (Modelo 3 – Misto B)

Na análise com recuperação de informação intergenotípica, apenas os efeitos das novas linhagens experimentais (progênies), além do erro experimental, são considerados de natureza aleatória. A estrutura de variabilidade das observações assume, portanto, dois componentes: a variância genética entre progênies (σ_g^2) e a variância do erro (σ_e^2). Como já referido, neste caso, os efeitos τ_i ($i=1,2,...,v$; $v=p+t$) são de duas naturezas: fixos, se *i* for uma testemunha, ou aleatórios se *i* for uma progênie. E, como os *softwares* estatísticos, a exemplo do *SAS*, ainda não estão estruturados para acomodar esse tipo de situação, as modelagens (I, II e III) usadas anteriormente não se prestam adequadamente para a presente análise.

A proposta de Scott & Milliken (modelo II) encontra apenas uma limitação de ordem computacional, pois, assumindo-se os efeitos $x_i(c_i)$ aleatórios ('**`random X(C);`**'), efeitos estes com variação nula dentro de testemunhas, teoricamente o modelo contemplaria a suposição de aleatoriedade dos efeitos de progênies. Contudo, na prática, entrando-se com este modelo no *PROC MIXED* do *SAS*, o programa calcula incorretamente (superestima) os erros padrão associados às médias de testemunhas, bem como os testes estatísticos relacionados à fonte de variação '**`C`**'. Isto

em virtude do conflito gerado ao admitir-se completa aleatoriedade de '**X(C)**', uma vez que o nível '*0*' de '**X**' corresponde às testemunhas, de natureza fixa (apesar da ausência de graus de liberdade dentro de cada testemunha). Os demais resultados são listados corretamente, incluindo-se as estimativas dos componentes de variância (σ^2_g e σ^2_e), os preditores dos efeitos aleatórios de progênies com seus respectivos erros padrão, bem como as estimativas das médias de testemunhas (exceto os erros padrão associados).

Diante desse fato, a alternativa apresentada por Wolfinger *et al*. (1997) é informar o modelo diferentemente ao *PROC MIXED*, através do uso de uma covariável. Para isso, propuseram uma variável binária (*dummy*) auxiliar, denominada '**NEW**', que indica se o tratamento (**GENOT**) é (**NEW=1**) ou não (**NEW=0**) um novo tratamento (progênie). A variável deve ser criada no arquivo de dados (**ARQ_ORIG**), no momento de sua leitura pelo *SAS*, conforme já apresentado (item *2.3.1*). Para a análise, contudo, esta variável não deve ser incluída como atributo de classificação no comando '**class**', mas, entra apenas na construção de uma covariável aleatória '**GENOT*NEW**' a ser introduzida no modelo e, portanto, listada no comando '**random**' do *PROC MIXED*. Logo, um conjunto de instruções que permite a realização da análise é (exemplo com *t=4* testemunhas):

```
proc mixed data=ARQ_ORIG;
  class BLOCO GENOT C;
  model PG=BLOCO C / ddfm=satterth chisq;
  random GENOT*NEW /solution cl;
    contrast 'TESTEMUNHAS'  C  0 1 -1 0 0,
                            C  0 0 1 -1 0,
                            C  0 0 0 1 -1;
    contrast 'TEST vs PROG' C -4 1 1 1  1;
  lsmeans C / cl;
  make 'solutionr' out=EBLUPs;
run;
```

A variável '**C**' ('*treatn*' de Wolfinger *et al*.,1997), como já definida, é igual a '**GENOT** ' (identificação do genótipo) para cada uma das testemunhas, mas, assume um nível constante (*0*) para todos os novos tratamentos (progênies). Sendo assim, é usada como um efeito fixo para modelar as médias de cada testemunha e uma média comum (μ_P) para todas as progênies. Estas, por sua vez, são assumidas variarem os seus efeitos aleatoriamente em torno da média μ_P, a qual é livre para variar em relação às médias das testemunhas (μ_1, μ_2, ..., μ_t). Assim, o comando '**model**' lista os efeitos fixos de '**BLOCOS**' e de '**C**', e sua opção '**ddfm=satterth**' possibilita o cálculo da aproximação de Satterthwaite para graus de liberdade associados às estatísticas *F*, χ^2 (**chisq**) e *t*, no caso de estas serem obtidas a partir de combinações lineares de quadrados médios (Piepho, 1997;

Duchateau & Janssen, 1997). Isto permite melhorar a eficiência dos testes aproximados, próprios de modelos lineares mistos sob desbalanceamento. As linhas associadas ao comando '**contrast**' (não obrigatórias) apenas determinam um desdobramento de interesse geral para a variação em '**C**'. E, ainda para os efeitos fixos, '**lsmeans**' calcula as estimativas das médias para os *t+1* níveis da variável '**C**': $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{\mu}_2$, ..., $\hat{\mu}_t$ (com $\hat{\mu}_0 = \hat{\mu}_P$).

O comando '**random**' lista como efeito aleatório apenas a nova fonte de variação '**GENOT*NEW**'. Esta variável recebe o valor *zero* para todas as testemunhas e um nível diferente para cada uma das progênies, a própria identificação do genótipo. A opção '**solution**', associada ao comando '**random**', requer, então, o cálculo dos preditores dos efeitos aleatórios (*EBLUP*'s) associados a cada uma das progênies, bem como os erros padrão associados. O termo *EBLUP* (*empirical best linear unbiased predictor*) é utilizado para indicar que a matriz **V** (variâncias-covariâncias das observações), usada na determinação do preditor, é estimada, ao invés de conhecida parametricamente (Littell *et al*., 1996; SAS Institute, 1997). Por fim, o comando '**make 'solutionr'** ...' cria um *dataset SAS*, aqui nomeado '**EBLUPs**', que armazena os *EBLUP*'s, seus respectivos erros padrão e intervalos de confiança. Estes preditores permitem ordenar e comparar as progênies entre si para fins de seleção (Wolfinger *et al.*, 1997), o que pode ser obtido com as instruções:

```
proc sort data=EBLUPs;
  by descending _est_;
proc print; run;
```
[1]

Havendo interesse nas médias ajustadas de progênies individuais (cada nível do fator aleatório '**GENOT*NEW**'), deve-se construir uma função linear combinando-se efeitos fixos e aleatórios: $\mathbf{w} = \mathbf{L}'\boldsymbol{\beta} + \boldsymbol{\gamma}$; em que: $\mathbf{L}'$ é a matriz de coeficientes dos efeitos fixos, na referida função; e, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são os vetores de efeitos fixos e aleatórios, respectivamente, no correspondente modelo linear misto $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$ (SAS Institute, 1997, p. 634). Searle *et al*. (1992) tratam deste problema comentando que, para $\mathbf{L}'\boldsymbol{\beta}$ estimável, o estimador ou preditor de **w** possui propriedades de *BLUP* (erro médio quadrático mínimo, linearidade em relação a **y** e não tendenciosidade), sendo, por isso, também denotado: *BLUP(***w***)*= $\tilde{\mathbf{w}} = \mathbf{L}'\boldsymbol{\beta}^0 + \tilde{\boldsymbol{\gamma}}$; com $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ e $\tilde{\boldsymbol{\gamma}} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$ (SAS

[1]/ Nas versões *SAS* inferiores à 6.12, as variáveis internas do *PROC MIXED*, do tipo '**_var_**', são referidas como '**var**', por exemplo: '**_est_**' como '**est**' ; '**_resid_**' como '**resid**'; etc.

Institute, 1997, p. 641). No presente caso, a função **w** de interesse é μ_P+g_i (sendo g_i o valor genotípico da progênie *i*), com preditor: $EBLUP(\mu_P+g_i) = \hat{\mu}_P + \tilde{g}_i$. Logo, $\mathbf{L}'\boldsymbol{\beta}^0 = \hat{\mu}_P$ é a estimativa da média comum de todas as progênies e $\tilde{g}_i$ é o *i-ésimo* elemento do vetor $\tilde{\gamma}$, o *EBLUP* associado à progênie *i* (*i=1,2,...,p*).

Para se obter esse conjunto de médias no *SAS*, é necessário entrar com uma linha de comandos para cada progênie, logo após a instrução '**random** ...'. A sintaxe adequada tem o seguinte formato (ex: médias das progênies 1 e 2, num ensaio com 2 testemunhas):

```
estimate 'Prog.01' intercept 1 C 1 0 0| GENOT*NEW 1;
estimate 'Prog.02' intercept 1 C 1 | GENOT*NEW 0 1;
```

Aqui, os coeficientes associados aos efeitos fixos, incluindo-se a constante μ (**intercept**), devem ser fornecidos antes da barra vertical '**|**' e os referentes aos efeitos aleatórios, após esta barra. Observe-se que a população de progênies corresponde ao primeiro nível do fator '**C**', de modo que é supérfluo listar os coeficientes (**0 0**) dos demais níveis (referentes às testemunhas). O mesmo é válido para o fator '**GENOT*NEW**', ou seja, apenas devem ser listados os coeficientes '**0**' de progênies que antecedem, alfabética ou numericamente, a de interesse ('**1**') na respectiva linha.

É notório que a construção dessas funções é uma tarefa trabalhosa, sobretudo se o número de progênies for elevado. Vale relembrar, contudo, que este trabalho é dispensável para fins de ordenamento dos genótipos, pois, pode-se chegar ao mesmo resultado com a ordenação obtida para os *EBLUP*'s dos efeitos genotípicos ($\tilde{g}_i$), haja vista que, neste caso, a estimativa $\hat{\mu}_P$ é constante para todas as progênies. Em casos de real necessidade das médias, outra opção é construir uma rotina que permita adicionar $\hat{\mu}_P$ a cada um dos valores $\tilde{g}_i$ armazenados no *dataset* '**EBLUPs**'.

Análise com recuperação das informações interblocos e intergenotípica (Modelo 4 – Misto C)

Para recuperar ambas as informações, interblocos e intergenotípica, os efeitos de blocos e de progênies devem ser de natureza aleatória, além do erro experimental. Por conseguinte, a estrutura de variabilidade das observações compreende três componentes de variância: σ_b^2, σ_g^2 e σ_e^2. Também neste caso, os efeitos τ_i são fixos, se *i* for uma testemunha, ou aleatórios se *i* for uma progênie, acarretando as dificuldades operacionais já mencionadas. Dessa forma, a presente análise deve ser implementada com base nas recomendações feitas para o *Modelo 3*. As instruções computacionais são, portanto, praticamente idênticas, a menos da mudança do efeito '**BLOCO**' da

linha de comandos '**model** ...' (de efeitos fixos) para a linha '**random** ...' (de efeitos aleatórios), conforme o programa a seguir (exemplo com *t=2* testemunhas):

```
proc mixed data=ARQ_ORIG;
  class BLOCO GENOT C;
  model PG= C / ddfm=satterth chisq;
  random BLOCO GENOT*NEW /solution cl;
    contrast 'TESTEMUNHAS'  C  0 1 -1;
    contrast 'TEST vs PROG' C -2 1  1;
  lsmeans C / cl;
  make 'solutionr' out=EBLUPs;
run;
```

Neste caso, o *dataset* '**EBLUPs**' armazenará, além dos preditores de progênies, os de blocos. Assim, antes de executar a rotina de ordenação das progênies (**proc sort data=**..., etc.) é conveniente extrair os preditores de blocos do referido arquivo. Isso pode ser feito com as instruções a seguir, atualizando-se o número de blocos exemplificado (**b=4**) na primeira linha de comandos:

```
let b=4;
data EBLUPs; set EBLUPs;
if _n_<(&b+1) then delete;
run;
```

Médias marginais e respostas percentuais em relação às testemunhas do bloco

Embora a obtenção de médias e a de porcentagens sejam operações fáceis, o tamanho dos experimentos exige, normalmente, alguma automação na prática destas tarefas. Assim, construiu-se também um programa *SAS*, listado no Apêndice 3.1, que permite obter, simultaneamente, as médias marginais (*Modelo 0*) e as produtividades percentuais em relação à media das testemunhas no bloco (*Modelo 5*). Como já mencionado, tais estatísticas foram utilizadas para ordenar e selecionar progênies, servindo-se como padrões comparativos em relação aos modelos estatisticamente mais elaborados (*Modelos 1, 2, 3* e *4*).

3. RESULTADOS E DISCUSSÃO

3.1. Avaliação dos diferentes modelos de análise estatística

3.1.1. Influência dos modelos na precisão das médias genotípicas

Os resultados da aplicação dos diferentes procedimentos de análise estatística, bem como das relações utilizadas para a comparação entre estes são apresentados nas Tabelas 3.2 a 3.7 (seção Anexos). Inicialmente, convém observar que: *i)* recupera-se a informação interblocos ao se passar da análise do *Modelo 1* para o *2* e do *Modelo 3* para o *4*; *ii)* recupera-se informação intergenotípica

passando-se da análise do *Modelo 1* para o *3* e do *Modelo 2* para o *4*; e *iii)* recuperam-se ambas as informações (interblocos e intergenotípica) passando-se da análise do *Modelo 1* para o *4*. Como os *Modelos 0* e *5* não permitem estimar componentes da variabilidade aleatória, estes não serão considerados nesta primeira avaliação.

A princípio, é conveniente inspecionar os resultados referentes às razões de estimativas de componentes de variância ($\phi_b = \hat{\sigma}_b^2 / \hat{\sigma}_e^2$ e $\phi_g = \hat{\sigma}_g^2 / \hat{\sigma}_e^2$) e aos valores de erro padrão médio para progênies ($\overline{s}_{\overline{Y}\mathrm{P}}$) e testemunhas ($\overline{s}_{\overline{Y}\mathrm{T}}$). Como já era esperado, na maior parte dos ensaios, os erros de médias de progênies foram superiores aos de testemunhas (Tabela 3.2), haja vista seu menor número de repetições. Todavia, numa análise recuperando informação intergenotípica (*Modelos 3* e *4*) e em experimentos cuja relação ϕ_g for muito baixa, como por exemplo AREÃO-26 (número 18 na listagem), ESALQ-34 (25) e ESALQ-101 (32), entre outros (19, 26 e 27), isto pode-se inverter. A razão deste fato decorre de $\overline{s}_{\overline{Y}\mathrm{T}}$ não envolver no seu cálculo a estimativa $\hat{\sigma}_g^2$, mas tão somente $\hat{\sigma}_e^2$; enquanto $\overline{s}_{\overline{Y}\mathrm{P}}$ leva em conta as duas estimativas, as quais ponderam inversamente a informação intergenotípica. Raciocínio idêntico explica a maior precisão associada aos contrastes de médias de tratamentos, numa análise que recupera a informação interblocos em relação à análise intrablocos; embora, neste caso, envolvendo as variâncias $\hat{\sigma}_b^2$ e $\hat{\sigma}_e^2$ (Federer & Wolfinger, 1998).

A relação ϕ_b, na maior parte dos ensaios, foi bastante baixa, com valores inferiores a *0,25* para *85%* deles, incluindo-se vários experimentos com valores nulos ou negativos (*−0,00*) para as estimativas $\hat{\sigma}_b^2$. Este fato não pode ser confundido com homogeneidade das respectivas áreas experimentais, haja vista os valores relativamente elevados de $\hat{\sigma}_e^2$. Mas indica, sim, uma pequena diferenciação entre os blocos, com heterogeneidade dentro deles; o que reflete, portanto, uma blocagem pouco efetiva nesse conjunto de ensaios. Tal constatação provavelmente decorre do fato de que a escolha de blocos nos ensaios de competição de genótipos, muitas vezes, tem sido apenas uma medida preventiva antes que uma estratificação criteriosa das unidades experimentais. Nestas condições, o uso de uma análise que permite recuperar a informação interblocos (contida no parâmetro σ_b^2), em geral, garante melhoria à qualidade das estimativas (ex: redução do erro padrão de médias de tratamentos). Na Tabela 3.2 é notório que os maiores valores de $\overline{s}_{\overline{Y}\mathrm{P}}$, quase sempre, estiveram associados à análise intrablocos (*Modelo 1*). Já a análise que recupera a informação interblocos (*Modelo 2*), na maior parte dos ensaios, proporcionou uma redução em torno de *10%* no erro de médias de progênies, em comparação ao modelo fixo ($\overline{s}_{\overline{Y}\mathrm{P}} / \overline{s}_{\overline{Y}\mathrm{P}}^{\mathrm{Fixo}}$). As exceções foram os

ensaios de número 10, 14, 16, 17 e 24 (ESALQ-13, ANHEMBI-21, ESALQ-24, AREÃO-25 e ESALQ-33, respectivamente), para os quais a blocagem mostrou-se efetiva e, por conseguinte, a análise intrablocos.

Quanto à razão ϕ_g, que tem relação direta com a herdabilidade de observações individuais ($h^2_{Y_{ij}} = \sigma^2_g/[\sigma^2_g + \sigma^2_e] = \phi_g/[1+\phi_g]$), nota-se uma variação bem maior, entre *0,01* a *16,06* (Tabela 3.2), o que indica herdabilidades bastante diferenciadas entre os ensaios. (As relações mais altas foram, inclusive, pouco realistas para o caráter produtividade de grãos). Analogamente à discussão anterior, para experimentos com valores de ϕ_g ou de $h^2_{Y_{ij}}$ elevados, como ESALQ-24 (16), AREÃO-25 (17) e ANHEMBI-31 (22), a análise intrablocos mostrou-se satisfatória, resultando em erros padrão de mesma magnitude aos da análise que recupera informação intergenotípica (*Modelo 3*). Vale observar que os valores elevados de ϕ_g, nestes experimentos, decorrem sobretudo da baixa magnitude de suas estimativas $\hat{\sigma}^2_e$, resultando também nos menores coeficientes de variação (entre *5%* e *9%*). De fato, nesta condição, a análise baseada em modelo misto com tratamentos aleatórios converge para a de modelo fixo (Robinson, 1991; Bueno Filho, 1997). Em outras palavras, sob herdabilidade elevada, não se justifica investir na sofisticação dos procedimentos seletivos, uma vez que métodos simples podem produzir resultados satisfatórios.

Por outro lado, na quase totalidade dos ensaios (*90%*), a análise baseada no *Modelo 3* produziu erros de médias de tratamentos bastante inferiores aos obtidos com a análise intrablocos. Nestes ensaios, a redução percentual em $\overline{s}_{\overline{Y}\mathrm{P}}$, resultante do uso da informação intergenotípica, nunca foi inferior a *20%* em comparação ao *Modelo 1* – fixo (última coluna da Tabela 3.2). Isto representa um ganho substancial na precisão das estimativas das médias de progênies, alvo principal da seleção. Demonstra-se, assim, a importância de se considerar a informação de que as progênies estão relacionadas entre si por uma origem comum. Vale esclarecer que não foi levado em conta, aqui, o parentesco genealógico das progênies, mas tão somente o fato de elas constituírem, juntas, o grupo dos novos tratamentos experimentais (população de progênies). Numa análise mais elaborada, as informações de genealogia ou de similaridade genética poderiam ser incorporadas para melhorar ainda mais a qualidade das predições. Também poder-se-ia refinar a análise subdividindo a população de progênies conforme os cruzamentos que lhes deram origem, o que implicaria na estimação de componentes de variância genotípica para cada subpopulação ou cruzamento (modelo e procedimentos computacionais descritos na seção *3.4*).

Ao contrário do que se poderia esperar, o uso concomitante dos dois tipos de informação (*Modelo 4*) não trouxe redução adicional no erro das médias de progênies em comparação ao *Modelo 3*. A redução relativa ao modelo fixo praticamente manteve-se no mesmo nível daquela produzida pela análise que recupera apenas a informação intergenotípica. Logo, as informações interblocos e intergenotípica não foram independentes para esse conjunto de ensaios. Isto talvez possa ser explicado pelo fato de os blocos, em boa parte dos ensaios, terem sido representados por progênies de uns poucos (1, 2 ou 3) cruzamentos comuns, antes que por uma amostra aleatória do conjunto total de progênies. Logo, os dois tipos de informação, neste conjunto de ensaios, podem ser inclusive redundantes. Assim, o *Modelo 4* não se mostrou parcimonioso em relação ao *Modelo 3*; embora, ainda assim, mais eficiente do que os *Modelos 1* e *2* em termos de assegurar maior precisão às médias estimadas de progênies.

Em síntese, a recuperação de informação inter-efeitos (Federer & Wolfinger, 1996; 1998) quase sempre trouxe algum benefício à precisão das estimativas, sobretudo quando se aproveitou a informação interprogênies. Observando-se a última coluna da Tabela 3.2, nota-se que os erros de médias de progênies resultantes dos *Modelos 2*, *3* e *4*, em mais de *90%* dos ensaios, foram inferiores aos correspondentes erros do *Modelo 1*. Com relação à precisão associada às médias de testemunhas, em geral, os modelos diferiram pouco entre si, haja vista tal categoria de tratamentos ter sido considerada de natureza fixa em todos os modelos. Reitera-se que as análises aproveitando informação inter-efeitos aqui implementadas, embora tenham computado covariâncias importantes, desconsideram outras cuja inclusão no modelo poderiam trazer melhorias adicionais à análise.

Deve-se ressaltar que os maiores benefícios associados aos modelos que usam a informação interprogênies estão relacionados às baixas herdabilidades, as quais resultam de dois fatores: *i)* reduzida variabilidade genotípica entre as progênies; e *ii)* elevada variabilidade local associada ao caráter produtividade de grãos, em cada experimento. No caso da soja, espécie com uma longa história de seleção artificial, as baixas estimativas de σ_g^2 podem estar relacionadas à estreita base genética do germoplasma avaliado (alta similaridade genética entre as linhagens experimentais de cada ensaio). Isto, apesar das medidas preventivas adotadas no Programa, contra a vulnerabilidade genética da soja cultivada (Vello, 1992). De outro lado, pode-se relacionar alguns fatores que contribuem para elevar as estimativas de σ_e^2 nos ensaios de avaliação genotípica: grande número de progênies, exigindo extensas áreas experimentais; pouca disponibilidade de sementes por material, determinando parcelas estreitas (1 ou 2 fileiras de plantas) e um baixo número de repetições (1 ou 2); e, uso de blocos grandes, com heterogeneidade dentro e pouca diferenciação entre si. Do ponto de vista prático, tratando-se de ensaios preliminares, relacionados

ao melhoramento de espécies como a soja, algumas destas características são inflexíveis. Sendo assim, a adoção de modelos de análise que permitam aproveitar melhor a informação gerada e disponível nos dados experimentais, torna-se fundamental para o sucesso dos programas de seleção.

3.1.2. Influência dos modelos na seleção dos genótipos

A Tabela 3.3 ilustra as correlações das médias genotípicas obtidas pelos diferentes procedimentos, incluindo-se, agora, os *Modelos 0* e *5* (médias marginais e médias percentuais em relação a testemunhas, respectivamente). Na maior parte dos ensaios, é notória uma alta correlação entre as médias produzidas pelos seis modelos, tanto em magnitude (correlação de Pearson) como em posicionamento (correlação de Spearman). Numa análise preliminar, isto sugere que, na maioria dos ensaios conduzidos na prática, estes métodos selecionariam praticamente os mesmos genótipos em termos de produtividade de grãos; pois, suas respectivas médias ordenam, individualmente, de maneira muito similar o conjunto dos genótipos avaliados.

Numa análise mais minuciosa observa-se, entretanto, que as correlações de Pearson envolvendo os *Modelos 3* e *4* caem sensivelmente quando a relação ϕ_g fica abaixo de *0,25* ($h^2_{Y_{ij}}$ *<0,20*), como exemplificam os ensaios: ESALQ-101 (32) e ESALQ-34 (25). Nestes casos, os valores absolutos das médias ajustadas por estes modelos, embora ainda fortemente correlacionados entre si, não se mantêm associados aos de médias ajustadas pelos *Modelos 1* e *2*, bem como às médias dos *Modelos 0* e *5* (os quatro últimos exibiram sempre médias altamente correlacionadas entre si). Conclui-se, portanto, que as altas correlações são determinadas por valores de herdabilidade, no mínimo, medianos (valores de σ^2_g não muito baixos e de σ^2_e não muito elevados). É oportuno informar que cerca de dez ensaios foram excluídos deste estudo exatamente por apresentarem elevados coeficientes de variação (acima de 35%), o que os enquadraria também no grupo com as baixas correlações. Assim, não seria prudente inferir que na maioria dos ensaios reais as correlações entre os métodos se manteriam elevadas.

Por outro lado, observa-se que as correlações de Spearman permaneceram altas, inclusive, para os dois ensaios anteriormente mencionados (Tabela 3.3). Isto indica que, embora as respostas médias estimadas, em kg/ha, diferiram substancialmente entre os dois grupos de modelos, (*0*, *1*, *2*, *5*) e (*3*, *4*), a ordenação destas médias ainda continuou muito similar entre os seis procedimentos. Logo, sob as condições experimentais representadas por esse conjunto de ensaios, de fato, deve haver uma grande concordância dos métodos aqui avaliados em termos de ordenamento dos genótipos. Entre estas condições deve-se ressaltar o balanceamento relativo mantido nos ensaios,

isto é, progênies quase sempre não repetidas e testemunhas aparecendo uma vez na maioria dos blocos.

A peculiaridade dos *Modelos 3* e *4* pode ser avaliada inspecionando-se o grau de dispersão (CV_P) do conjunto de médias produzidas pelos diferentes procedimentos (Tabela 3.4). Observa-se que, na maioria dos ensaios, os dois modelos (com recuperação de informação interprogênies) tendem a produzir médias genotípicas mais uniformemente distribuídas do que os outros, sobretudo nos ensaios com herdabilidade básica ($h^2_{Y_{ij}}$) reduzida. Trata-se do efeito *shrinkage* ("encolhimento") das médias preditas em comparação às médias ajustadas por modelos de efeitos fixos, decorrente de uma menor variância do fator aleatório em relação à variância do erro (Robinson, 1991; SAS Institute, 1996). Esta propriedade das médias *BLUP* (ajustadas por modelos mistos com tratamentos aleatórios) perde importância prática à medida que a herdabilidade aproxima-se do limite máximo $h^2_{Y_{ij}} = 1$ (condição assumida pela análise de modelo fixo). Nota-se, por exemplo, pequena ou quase nenhuma convergência das médias *BLUP* em relação aos demais métodos, nos ensaios ANHEMBI-15 (7), ESALQ-24 (16) e ANHEMBI-31 (22), cujas herdabilidades foram as maiores. Por outro lado, no ensaio ESALQ-101 (32), as médias de progênies ajustadas pelos *Modelos 3* e *4* foram praticamente idênticas entre si (em cada modelo), convergindo para a média geral de progênies (a média da população da qual se originaram: $\hat{\mu}_P$). Isto, em função da estimativa quase nula da variabilidade genotípica relativa ($\phi_g = \hat{\sigma}^2_g / \hat{\sigma}^2_e = 0,02 \Leftrightarrow h^2_{Y_{ij}} \cong 0,02$).

É necessário esclarecer que, embora coeficientes de correlação sejam normalmente utilizados nesse tipo de estudo, estas medidas são mais fortemente determinadas pelo ordenamento dos valores das variáveis do que por diferenças (em magnitude) nestes valores. Assim, convém inspecionar outros critérios de avaliação comparativa. Mesmo porque critérios seletivos outros podem ser considerados num programa de melhoramento, por exemplo: *i)* seleção das progênies que superam a *performance* média da melhor testemunha (padrão comercial); *ii)* seleção das progênies com rendimento superior à produtividade média regional; etc. Nestes casos, métodos que ordenam igualmente os genótipos podem produzir seleções distintas em função das diferenças nas respectivas médias estimadas.

Os resultados da Tabela 3.5 mostram que, se apenas as linhagens superiores à melhor testemunha (em produtividade de grãos) forem selecionadas, os seis métodos de obtenção das médias genotípicas podem produzir seleções de intensidades bastante diferenciadas. Observa-se que, na maioria dos experimentos, os *Modelos 3* e *4* tenderam a selecionar um menor número de

linhagens do que os outros modelos (*0*, *1*, *2* e *5*). Estes, por sua vez, resultaram em proporções de genótipos selecionados muito similares entre si, em toda a série de ensaios. Considerando-se, adicionalmente, as altas correlações entre suas estimativas de médias, bem como entre os respectivos postos (Tabela 3.4), pode-se concluir que os quatro métodos, em geral, levariam a seleções praticamente idênticas. Esta elevada concordância, todavia, pode ter sido aumentada em decorrência da reduzida variabilidade entre blocos, incidente nesse grupo de experimentos.

Quanto aos *Modelos 3* e *4*, na maior parte dos ensaios, a proporção de genótipos retida para o próximo ciclo seletivo reduziu-se sensivelmente em relação aos outros modelos, quando a herdabilidade e a média da população de progênies foram baixas, isto é, $h^2_{Y_{ij}} < 0,5$ e $\hat{\mu}_P < \hat{\mu}_{(T.sup)}$ (Tabela 3.5). Observa-se, por exemplo, que nos ensaios 1 e 2 (AREÃO-11 e AREÃO-12), enquanto os *Modelos 0*, *1*, *2* e *5* retêm cerca de *30%* dos genótipos, os *Modelos 3* e *4* preservam, no máximo, *18%* deles. Em situações extremas, como exemplifica o ensaio AREÃO-26 (18), os dois modelos chegam a apontar para a interrupção do programa de seleção (nenhum genótipo selecionado); enquanto os demais continuam retendo *23%* dos genótipos. Outros ensaios (10, 14, 19, 25, 26, 27, 31 e 32) ilustram a mesma situação, embora com menores proporções de genótipos para os modelos *0*, *1*, *2* e *5*. Esta constatação tem um forte impacto em aplicações, haja vista o comprometimento de recursos, quase sempre limitados, para a condução de um programa de melhoramento.

Admitindo-se, hipoteticamente, um experimento em que a modelagem ideal fosse representada pelos *Modelos 3* ou *4* (modelos com tratamentos adicionais de efeitos aleatórios), uma seleção baseada em médias obtidas a partir dos *Modelos 0*, *1*, *2* e *5* poderia implicar num considerável desperdício de tempo e recursos (área, adubação, plantio, cultivo, coleta de dados, etc.). Além disso, a manutenção de linhagens pouco promissoras na etapa seguinte do processo compete com uma boa avaliação dos genótipos superiores, isto é, prejudica a precisão das estimativas para os genótipos de real interesse. Isto demonstra a importância da especificação do modelo a ser adotado na análise estatística, cujos resultados darão suporte à seleção dos genótipos. Confirma-se também a preocupação de Bueno Filho (1997) com respeito ao costumeiro tratamento de efeitos aleatórios como fixos, nos processos de estimação/predição de valores genotípicos. O autor adverte que isto tem levado a erros na seleção de genótipos, o que pôde ser aqui evidenciado.

Uma análise mais específica é possível quando se classifica cada população de progênies (cada ensaio) em: *i)* população de alto potencial produtivo, se $\hat{\mu}_P \geq \hat{\mu}_{(T.sup)}$; e *ii)* população de baixo potencial produtivo, se $\hat{\mu}_P < \hat{\mu}_{(T.sup)}$. Assim, considerando-se ainda a seleção das progênies superiores à melhor testemunha, constata-se que os *Modelos 3* e *4* são menos rigorosos do que os

demais se as populações são do primeiro tipo e a variabilidade genotípica relativa (herdabilidade) é baixa. Os ensaios ANHEMBI-92i (28), ANHEMBI-23 (12) e AREÃO-13 (5) exemplificam a situação (Tabela 3.5). Estes casos, embora menos freqüentes (população nova com média superior ao melhor padrão comercial), indicam que os *Modelos 3* e *4*, nesse tipo de seleção, garantem maiores oportunidades aos genótipos medianos destas populações, sobretudo quando a precisão dos ensaios for baixa (valores relativamente elevados de σ_e^2). Por outro lado, os dois modelos descartam mais intensamente genótipos medianos se estes estiverem relacionados a populações de baixo potencial, especialmente quando a herdabilidade é baixa. Esta maior sensibilidade dos *Modelos 3* e *4* à redução na herdabilidade pode ser ratificada comparando-se ensaios repetidos (ex: AREÃO-11/AREÃO-12, ESALQ-11/ESALQ-12, AREÃO-13/AREÃO-14, ANHEMBI-23/ANHEMBI-24, etc.). Observa-se, portanto, que os modelos que desconsideram a interdependência dos genótipos na estrutura de variabilidade dos dados, são muito menos flexíveis à variação em $h^2_{Y_{ij}}$ (Tabela 3.5).

3.1.3. Um caso ilustrativo

Para ilustrar as relações entre o conjunto de médias obtidas pelos diferentes procedimentos, escolheu-se o ensaio AREÃO-12, conduzido em 1992/93. O ensaio foi implantado para avaliar 1260 progênies (Tabela 3.1), distribuídas em 28 blocos de tamanho k=55 parcelas (45 progênies + 10 testemunhas). Em decorrência de observações perdidas ou discrepantes, das 1540 parcelas originais foram aproveitadas 1379 observações para o caráter produtividade, perfazendo a avaliação de 1228 progênies (linhagens experimentais).

Considerando-se o uso comum da seleção com base em testemunhas intercalares (*Modelo 5*) ilustra-se, aqui, o comportamento das *40* progênies melhor classificadas por este método, em relação aos outros procedimentos (Tabela 3.6). Observa-se que, embora estas progênies, na sua maioria, tenham tido boas classificações também pelos *Modelos 0*, *1*, *2*, *3* e *4*, algumas delas teriam grande chance de serem descartadas por um ou outro destes modelos. Considere-se, em princípio, uma seleção truncada de *10%* dos novos genótipos superiores, ou seja, retenção das 123 linhagens mais produtivas para o próximo ciclo seletivo. Nota-se que, em relação ao *Modelo 0*, catorze das *40* linhagens seriam descartadas. Entre estas estão: USP-16638, USP-11589, USP-16180 e USP-13993, todas classificadas entre as dez primeiras pelo *Modelo 5*, mas, acima da 123ª posição pelo *Modelo 0*. Já em relação à análise intrablocos (*Modelo 1*), apenas a linhagem USP-14342 (39) seria descartada. Isto mostra que a seleção com base em testemunhas, aqui avaliada, tem grande concordância com a seleção produzida pela análise de modelo fixo. De fato, no delineamento de blocos aumentados a análise intrablocos promove o ajuste de tratamentos, quase sempre, baseando-

se na média intrablocos das testemunhas. Exceções ocorrem, por exemplo, quando alguns novos tratamentos são repetidos entre os blocos, o que faz com que as informações destes tratamentos também contribuam para o referido ajuste.

Com a recuperação da informação interblocos (*Modelo 2*) a discordância em relação ao *Modelo 5* volta a aumentar, à semelhança daquela com médias marginais. Nesse caso, onze das *40* linhagens seriam descartadas (Tabela 3.6). A quase coincidência deste descarte com o do *Modelo 0* reflete a tendência da análise interblocos para o modelo de médias simples não ajustadas ($Y_{ij}=\mu+\tau_i+e_{ij}$), quando a variabilidade entre blocos for baixa (ϕ_b=0,27). De fato, sob $\sigma_b^2 = 0$ as duas análises coincidem, produzindo médias de tratamentos idênticas, ou seja, sem qualquer ajuste para os efeitos de blocos (Scott & Milliken, 1993).

Fazendo-se a mesma análise em relação aos *Modelos 3* e *4*, observa-se que oito das *40* linhagens seriam descartadas, haja vista ocuparem más posições de classificação (acima da 123ª): USP-13442 (12), USP-13957 (13), USP-17160 (27), USP-16226 (28), USP-172246 (29), USP-17654 (32), USP-16776 (34) e USP-14342 (39). A coincidência perfeita dos genótipos descartados nos dois modelos e a similaridade de postos da maioria das linhagens (Tabela 3.6) reforçam a semelhança destes modelos, em termos de seleção, já comentada anteriormente. Nota-se ainda que, pelos *Modelos 3* e *4* cerca de *140* linhagens superaram a melhor testemunha (Hale-321), enquanto pelos outros modelos este número nunca foi inferior a *360*. Federer (1998) obteve resultados similares para um delineamento aumentado de linhas e colunas: *52* novos genótipos acima da testemunha superior na análise intrablocos, contra *36* na análise com recuperação da informação intergenotípica. Isso demonstra o que já se discutiu sobre o maior rigor dos *Modelos 3* e *4*, quando a média da população é inferior ao limite de descarte ($\hat{\mu}_{(T.sup)}$) e a herdabilidade é baixa ($h^2_{Y_{ij}}$ *=0,37*).

A Tabela 3.7 apresenta a listagem das *123* melhores linhagens (*10%* superiores) para cada um dos procedimentos estatísticos. No final desta Tabela são ainda apresentadas as porcentagens de coincidência de genótipos selecionados entre os diferentes modelos. Os valores indicam que a seleção com base em testemunha (*Modelo 5*) mostrou a menor concordância geral com as demais, exceto com a do *Modelo 1*. Os resultados reforçam as constatações anteriores, revelando também as concordâncias entre as seleções produzidas pelos *Modelos*: *0* e *2*; *1* e *2*; e, *3* e *4*.

Algumas relações entre as médias genotípicas ajustadas pelos modelos estatisticamente mais elaborados (*Modelos 1* a *4*), podem ainda ser visualizadas nas Figuras 3.1 a 3.6 (Anexos). Para fins de ilustração, tomou-se apenas uma amostra aleatória de dez das *1228* progênies, mais as duas testemunhas melhor classificadas. De uma maneira geral, as Figuras permitem evidenciar: *i)* a

possibilidade de trocas de posições relativas das linhagens entre os diferentes modelos (diferentes linhagens liderando a classificação); *ii)* a tendência dos *Modelos 3* e *4* produzirem médias mais uniformemente distribuídas do que os outros modelos (efeito *shrinkage* sobre as médias de progênies, não sobre as de testemunhas); *iii)* os diferentes números de linhagens superando as melhores testemunhas, sobretudo quando se recupera (*Modelos 3* e *4*) ou não (*Modelos 1* e *2*) informação intergenotípica; e, *iv)* a concordância dos *Modelos 3* e *4* (talvez em decorrência de efeitos de blocos de pequena intensidade).

Em síntese, essas avaliações permitem concluir que o uso de diferentes métodos para a obtenção das médias genotípicas pode levar a seleções diferenciadas. E, a despeito da concordância aparentemente alta entre os procedimentos avaliados, genótipos bem classificados por alguns deles podem ser descartados por outros. Considerando-se que a expectativa normal de um ciclo seletivo, numa espécie já bastante melhorada como a soja, é a liberação de um ou dois cultivares, esta diferenciação pode determinar o maior ou menor êxito do programa.

3.2. Aspectos experimentais relacionados à análise estatística

O estudo teórico-prático implementado neste trabalho permitiu obter algumas orientações experimentais para questionamentos freqüentes relacionados à análise estatística de delineamentos aumentados. Enfocar-se-á, inicialmente, a questão dos tratamentos testemunhas. Quanto à sua escolha, a recomendação da literatura é que se use pelo menos dois cultivares de base genética similar ao material de teste. Kempton & Gleeson (1997) reportam uma pesquisa em que o uso de progenitores para o ajustamento de respostas genotípicas resultou em menor variação local do que se baseado em outros cultivares. Segundo Bearzoti (1994), em blocos aumentados, é fundamental que as testemunhas representem bem a variância residual da população segregante. Isto justifica-se porque, na maior parte dos casos (ensaios com tratamentos adicionais não repetidos), o erro experimental (σ_e^2) é estimado como a variância média dentro de testemunhas, após o ajuste para blocos. De fato, as observações dos novos tratamentos, repetidos uma só vez, apenas contribuem para as estimativas deles próprios, nada informando a respeito de blocos, média geral, ou *erro* (Federer, 1998).

Neste sentido, uma recomendação pouco difundida, mas importante para a melhoria da precisão experimental, é que se procure repetir, em blocos diferentes, os novos tratamentos que disponham de maior quantidade de sementes, a despeito de isto aumentar o desbalanceamento. Esta prática implicará num aumento no número de graus de liberdade para estimar a variância do *erro*, resultando em maior precisão experimental e numa melhor representatividade desta estimativa. Isso

porque, neste caso, a estimativa resultará de uma combinação de informações das testemunhas e destes novos tratamentos. Os benefícios estendem-se também à estimação da média geral, aos ajustes para blocos e à precisão das estimativas de médias dos referidos tratamentos adicionais.

Outro aspecto experimental importante refere-se à freqüência de parcelas com variedades testemunhas. Em geral, aumentando-se o número destas parcelas tem-se uma melhoria no controle da variação local. Contudo, para um número fixo de parcelas, isto implica numa redução do número de genótipos avaliados. Assim, Kempton & Gleeson (1997) não recomendam o uso de uma alta freqüência de parcelas testemunhas, por exemplo, acima de uma em cinco parcelas; a menos que haja indícios de forte heterogeneidade espacial de pequeno alcance (quando a correlação entre observações de parcelas vizinhas cai rapidamente com o aumento da distância entre elas). Por outro lado, um número reduzido destes tratamentos traz prejuízos imediatos à precisão experimental, sobretudo se o número de blocos for baixo. Assim, o balanço prático que os melhoristas têm encontrado situa-se entre *dois* e *quatro* cultivares.

É oportuno observar que, em blocos aumentados, um número elevado de observações, por si só, não é suficiente para caracterizar um experimento como grande. Um ensaio com, por exemplo, *4* blocos, *3* testemunhas (repetidas uma vez em cada bloco) e *220* progênies (repetidas uma só vez no experimento), apesar de possuir *232* unidades experimentais e um número de testemunhas dentro da faixa recomendável, não pode ser considerado grande. Isto pois, em sua análise de variância apenas *seis* graus de liberdade são disponíveis para a estimativa da variância residual. Considerando-se que o número de testemunhas necessariamente será baixo (entre 2 e 4), para se garantir um experimento mínimo (com $GL_{Erro} \geq 10$) seriam necessários pelo menos *cinco* blocos. Mas, se o número de testemunhas cair para *dois*, o de blocos nunca deverá ser inferior a *dez*.

Um aspecto polêmico, mas pouco comentado nos textos básicos sobre delineamentos em blocos, é o relativo à interação "tratamentos *x* blocos". No caso de blocos aumentados, Bearzoti (1994) manifesta uma preocupação com o fato da variância do erro ser obtida a partir da interação "tratamentos comuns *x* blocos". A despeito da validade desta preocupação, deve-se esclarecer que uma forte suposição da *ANOVA* clássica dos delineamentos em blocos é a de aditividade dos efeitos de blocos e de tratamentos, ou seja, ausência de interação "tratamentos *x* blocos". Com efeito, na sua presença, até mesmo uma recomendação geral de tratamentos ficaria impossibilitada. Mas, uma vez atendida a suposição, o quadrado médio residual mensura apenas o erro experimental, isto é, a variação média decorrente de causas intrínsecas às unidades experimentais (Helms *et al*., 1999). Por outro lado, na prática, a interação pode existir e, ignorá-la por conveniência analítica certamente introduzirá prejuízos à qualidade das conclusões obtidas. De qualquer forma, essa limitação

estatístico-experimental, embora explícita nos delineamentos aumentados, constitui uma deficiência geral dos delineamentos em blocos, a qual carece ainda de maior investigação.

Gusmão (1986) reporta que, suposições como esta, bem como a de homogeneidade dentro de blocos, são violadas em boa parte das aplicações destes delineamentos. Stroup & Mulitze (1991) acrescentam que, apesar da fé quase religiosa de muitos pesquisadores no delineamento de blocos completos casualizados (BCC), na presença de variabilidade sistemática intrabloco, sua análise tradicional pode ser tão pobre quanto catastrófica. Mariotti *et al*. (1997) tiveram constatações semelhantes, particularmente em experimentos com mais de quinze tratamentos. Baseado em fatos dessa natureza, relacionados sobretudo com as suposições subjacentes à análise, Gusmão (1986) tenta mostrar que o delineamento de BCC é inadequado para testes de cultivares.

Os grandes progressos obtidos historicamente no melhoramento genético, resultantes da utilização freqüente desses delineamentos, parecem refutar tão contundentes afirmações. Isto, entretanto, não exime os aplicadores de uma busca constante de instrumentos diagnósticos para verificar a validade das suposições, bem como de delineamentos e métodos menos restritivos de análise. Diante disso, uma recomendação no sentido de permitir avaliar a referida interação, em delineamentos aumentados, seria repetir as testemunhas também dentro de blocos. Isto confere, em favor da precisão experimental, mais graus de liberdade à variância do *erro* do que o aumento no número de blocos (Bearzoti, 1994). Ademais, possibilita o teste dessa interação contra o *erro puro* (variância entre parcelas igualmente tratadas dentro de blocos), o que é inclusive um indicativo da necessidade ou não de se aplicar algum tipo de análise espacial (Helms *et al*., 1999).

3.3. Sobre as suposições para os efeitos de blocos e de progênies

Na seção *3.1* apontaram-se possíveis diferenças entre os métodos de análise estatística aqui avaliados, com respeito ao resultado final da seleção. Tomando-se os modelos estatisticamente mais elaborados, observou-se influência marcante da suposição associada aos efeitos de progênies (tratamentos novos), se fixos ou aleatórios. É mister, entretanto, estender um pouco mais a discussão acerca das suposições destes modelos, com vistas a orientar o usuário para uma especificação adequada do modelo da análise estatística.

Inicialmente, é oportuno mencionar que o modelo fixo aqui representado (*Modelo 1*) corresponde à proposta original de Federer (1956). Nesta proposta, o interesse estatístico está voltado para a obtenção de uma análise de variância que permita obter estimativas da média geral, das médias de tratamentos ajustadas para os efeitos de blocos, de contrastes entre médias de tratamentos e das variâncias associadas ao erro experimental e aos contrastes de interesse. Por esta

abordagem assume-se uma estrutura simples para as matrizes de variâncias-covariâncias dos erros e das observações ($\mathbf{R}=\mathbf{I}\,\sigma_e^2 \Rightarrow \mathbf{V}=\mathbf{I}\,\sigma_e^2$). Isto significa que se admite independência completa entre as observações ou unidades experimentais. Para garantir pequenos desvios desta suposição, Federer (1956; 1961a) e Federer & Raghavarao (1975) recomendam que as testemunhas sejam alocadas ao acaso dentro dos blocos e, somente em seguida, distribuindo-se os tratamentos adicionais, também por sorteio, às parcelas restantes de todo o experimento. Os autores acrescentam que, se alguns destes tratamentos dispuserem de material de propagação suficiente para mais de uma repetição, estas devem ser feitas em blocos diferentes de modo a garantir maior precisão experimental.

Diante disso, pode-se afirmar que alocações sistemáticas de testemunhas, em parcelas de posições prefixadas (normalmente eqüidistantes), viola o requisito básico introduzido para se tentar garantir independência entre as observações, a casualização. Neste sentido, todos os modelos apresentados no presente trabalho (incluindo-se os da seção *3.4*) assumem igual estrutura para a matriz de erros ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$), embora com matrizes $\mathbf{V}$ diferenciadas. Isto porque, em nenhum desses modelos se propôs estimar alguma dependência espacial entre parcelas vizinhas para incorporá-la à estimação/predição e testes de hipóteses. Dessa forma, todos estão assentados sobre as premissas básicas de Federer (1956), com respeito a alocação dos tratamentos às parcelas.

A rigor, uma distribuição sistemática implicaria numa de duas alternativas: ou o melhorista contenta-se com métodos de ajustamento de respostas em função das observações nas testemunhas (com reduzidas possibilidades de inferência estatística); ou ele avalia a correlação espacial incidente no experimento e, no caso de presença significativa, faz uso de procedimentos estatísticos menos restritivos, que permitam lidar com erros correlacionados (matriz $\mathbf{R}$ não diagonal). Neste último caso (o mais recomendado), buscar-se-ão métodos estatísticas que minimizem os efeitos desse tipo de autocorrelação na seleção de tratamentos e nas estimativas de parâmetros como componentes de variâncias, herdabilidade, etc. Embora não tratados neste artigo e ainda pouco divulgados, alguns métodos já estão disponíveis para se aplicar este tipo de abordagem aos ensaios genéticos com tratamentos não repetidos (Cullis *et al*., 1989; Federer, 1998; Cullis *et al*., 1998).

No que se refere às análises com recuperação de informação interblocos (*Modelos 2* e *4*) é mister informar que, para a validade das estimativas combinadas dos efeitos no modelo, os conjuntos de tratamentos que representam os blocos devem ser constituídos de forma aleatória (Rao, 1947; Malheiros, 1982). Logo, esse tipo de análise não admite que os blocos correspondam a conjuntos especiais de tratamentos, ou seja, cada bloco não pode, preferencialmente, ser formado por tratamentos com uma mesma procedência (ex: progênies vindas de um mesmo cruzamento).

Ademais, nestas situações, os efeitos de blocos acabam confundidos com os efeitos de procedência, o que descaracteriza sua natureza aleatória e impossibilita a sua estimação imparcial. Este aspecto da aleatoriedade de blocos, em geral, é assegurado nos ensaios de teste de progênies em que estas são oriundas de uma só população. É comum, porém, nas situações em que as progênies têm procedências diversas, uma avaliação dos genótipos em blocos ou conjuntos (*set* em inglês) que têm alguma amarração à origem do material. Nestes casos, a rigor, o analista fica impedido de desfrutar das vantagens que a recuperação da informação interblocos propicia à análise estatística.

Alguns autores entendem que a aleatoriedade dos efeitos de blocos só pode ser assumida se, de fato, estes representarem uma amostra aleatória de uma população de blocos (Gusmão, 1986; Piepho, 1994). E, a forma usual como os blocos têm sido selecionados nos campos experimentais, na maioria das vezes, não oportuniza qualquer seleção alternativa de outro conjunto de blocos. Dessa forma, para dar validade à referida suposição, dever-se-ia evitar o estabelecimento de blocos lado a lado na área experimental, mas, amostrá-los na área disponível. Dado que isto é praticamente inexeqüível, estes autores argumentam que os efeitos de blocos, em geral, devem ser assumidos como fixos, tal como nos *Modelos 1* e *3*.

Outros autores, contudo, são mais flexíveis nesse tipo de definição, entendendo que assumir um fator como de efeitos aleatórios, quase sempre, é uma prática salutar; sobretudo se o número de níveis do fator em consideração for elevado. Assim, compreendem que, se não houver interesse em manter os mesmos níveis (blocos) numa possível repetição do experimento, os correspondentes efeitos têm caráter aleatório (Jiménez & Villa, 1995). Isso, sem dúvida, é verdade para os efeitos de blocos, na maioria dos experimentos; o que justificaria, portanto, a adoção da chamada análise combinada, incluindo a recuperação de informação interblocos. Federer & Wolfinger (1998) reforçam essa tese, afirmando que os blocos de um experimento não têm importância outra senão a forma como afetam as médias de tratamentos, sendo, por isso, realisticamente aleatórios. Neste caso, as correspondentes médias ajustadas de tratamentos são menos afetadas pela natureza dos blocos incompletos, que têm, com a recuperação de informação interblocos, seus efeitos reduzidos (Federer & Wolfinger, 1998).

Sob este ponto de vista, a exigência básica no sentido da estimação de σ_b^2 e da eficiência da análise recuperando informação interblocos, é a de que o número de blocos não seja inferior a *dez* (Malheiros, 1982). Uma recomendação que, por certo, pode ser acrescida é a do uso de blocos de pequeno tamanho. Acredita-se, entretanto, que a afirmação daqueles autores de que é sempre desejável modelar os efeitos de blocos como aleatórios esteja assentada, obrigatoriamente, sob as premissas de casualização dos delineamentos clássicos, conforme definiu Federer (1956; 1961a).

Assim, no caso de blocos aumentados, continua sendo única a restrição à casualização completa (ainda sob suposição de independência): a de blocos em relação aos tratamentos repetidos (testemunhas e progênies com mais de uma parcela). Sob tais condições, sim, o pesquisador poderá usufruir integralmente dos benefícios estatísticos conferidos pela análise com recuperação da informação interblocos.

Sobre as suposições relacionadas aos efeitos de tratamentos, se fixos ou aleatórios, cabem algumas considerações preliminares. Embora ainda pouco difundida entre os melhoristas de plantas, a metodologia de modelos lineares mistos, incluindo a predição de efeitos aleatórios de tratamentos, foi desenvolvida por Charles R. Henderson em 1948 (Henderson, 1984). A princípio (até início dos anos oitenta), as limitações de recursos computacionais restringiram fortemente a sua divulgação. Historicamente, então, os modelos mistos têm sido analisados através de procedimentos apropriados para modelos fixos (Montebelo, 1997). E, os resultados relativamente satisfatórios deste enfoque contribuíram para retardar ainda mais a sua aplicação no melhoramento de plantas. Tais simplificações, entretanto, em situações de desbalanceamento ou de heterogeneidade de variâncias, levam à perda de precisão e de valor explicativo, podendo implicar em erros na seleção dos genótipos (Bueno Filho, 1997).

Segundo Stroup & Mulitze (1991), os estatísticos práticos tendem a distinguir entre efeitos fixos e aleatórios, mesmo num modelo misto, conforme as abordagens tradicionais de modelos fixos e aleatórios; isto é, comumente são priorizadas duas formas clássicas de estimação, o *BLUE* (*best linear unbiased estimator*) de funções estimáveis dos efeitos fixos e os componentes de variância associados aos efeitos aleatórios. Todavia, em modelos lineares mistos é disponível uma terceira forma de inferência. Trata-se da predição de valores realizados de variáveis aleatórias ou de funções dos efeitos aleatórios, isoladamente ou em combinação com efeitos fixos, o chamado *BLUP* (*best linear unbiased predictor*). Embora os componentes de variância sejam importantes, estes não refletem toda a "história" subjacente aos efeitos aleatórios, já que se pode ter interesse também em estimar os efeitos aleatórios individuais (Verbeke & Molenberghs, 1997). Em outras palavras, o fato de considerar genótipos como aleatórios, ao invés de fixos, não necessariamente elimina o interesse nas respostas de genótipos específicos (Piepho, 1994).

Nos modelos de blocos aumentados, a controvérsia relacionada à suposição dos efeitos de tratamentos, em geral, restringe-se aos novos tratamentos (progênies), pois as testemunhas, na maioria dos casos, são inquestionavelmente de efeitos fixos. A suposição de tratamentos de efeitos fixos (*Modelos 1* e 2), em princípio, é natural dos ensaios finais do processo de avaliação. Isto justifica-se por que os novos genótipos caracterizam populações individualmente constituídas, tal

como as testemunhas. Mas, ainda assim, ignorar o relacionamento entre as novas linhagens experimentais, pressupondo fixos seus efeitos, pode significar alguma perda de informação, a chamada *informação intergenotípica* ou *intervarietal* (Federer & Wolfinger, 1998; Federer, 1998). Por isso, vários autores entendem que, admitir os efeitos genotípicos como aleatórios (*Modelos 3* e *4*) é, quase sempre, uma prática salutar (Hill & Rosenberger, 1985; Stroup & Mulitze, 1991; Piepho, 1994; Bueno Filho, 1997; Federer & Wolfinger, 1998; Federer, 1998).

Stroup & Mulitze (1991) afirmam que, se o número de níveis de um fator (ex: tratamentos) for grande (algo entre 20 e 100) e a distribuição de seus efeitos for razoavelmente simétrica, *BLUP*'s são tipicamente mais eficientes do que *BLUE*'s. Assim, com freqüência, é preferível modelar seus efeitos como aleatórios, a despeito das definições tradicionais classificá-los como fixos, e isso parece ser particularmente verdadeiro com dados desbalanceados. Os autores concluem com uma afirmação contundente: "a distinção tradicional entre efeitos fixos e aleatórios não é útil e pode, de fato, levar o analista a escolher uma alternativa menos eficiente". Piepho (1994) comunga do mesmo ponto de vista. Federer (1998) advoga que toda informação presente num experimento (interblocos, interlinhas, intercolunas, intergradientes, intergenotípica) deve ser extraída, exemplificando a classe dos delineamentos aumentados como uma situação em que uma parte ou todos os efeitos genotípicos podem ser considerados aleatórios. Comenta ainda que, ignorar esse tipo de informação é como ignorar a informação de parcela num delineamento de parcelas subdivididas. Enfim, recuperando-se as informações associadas aos efeitos aleatórios, em geral, obtém-se análises mais eficientes (com menores erros para as estimativas) e, por conseguinte, tem-se uma melhor utilização dos recursos experimentais (Wolfinger *et al*., 1997).

Muito deste pensamento assenta-se sobre o argumento de que o uso dessas informações corresponde à adoção de uma modelagem menos restritiva, a qual converge naturalmente para a mais restrita (de efeitos fixos) se a estrutura dos dados assim se expressar. McLean *et al.* (1991) enfatizam que os procedimentos de modelos mistos podem ser aplicados a qualquer modelo, balanceado ou não, e com uma estrutura geral de covariâncias; ou seja, os procedimentos de análise de modelos fixos e de modelos aleatórios são casos particulares (extremos) da metodologia de modelos mistos. Por isso, a sua aplicação leva a uma abordagem unificada para modelos lineares. A despeito disso, entende-se que, assumir um fator como de efeitos aleatórios, ao invés de fixos, implica numa ampliação do espaço de inferência (SAS Institute, 1996). E, na ausência da *população conceitual* a que se refere Henderson (1984), isto parece não se justificar.

Diante dessa retrospectiva teórica, pode-se concluir que a suposição de aleatoriedade para os novos tratamentos (*Modelos 3* e *4*) não se limita, necessariamente, aos ensaios relacionados às

fases preliminares do processo seletivo. Mas fica na dependência, sobretudo, do número de genótipos avaliados (algo superior a 20) e da existência de algum relacionamento entre estes, que justifique a estimação do componente de variância σ_g^2.

O uso da informação de parentes para melhorar a predição de valores genotípicos individuais, sobretudo quando se dispõe de pouca ou nenhuma informação individual, representa um grande trunfo da modelagem mista com genótipos aleatórios. Por isso, mesmo numa população mais restrita, se o parentesco entre os genótipos realmente existir, ignorá-lo supondo-os de efeitos fixos significa perda de informação. Isto equivale a predizer a média genotípica utilizando-se apenas dados diretos (das repetições do genótipo), quando os dados indiretos (dos outros genótipos) também informam sobre o comportamento do genótipo sob predição (Gauch, 1992). Segundo Bueno Filho (1997), a inclusão de parentes garante maior conexão entre diferentes experimentos, o que permite dispensar o uso de testemunhas e, inclusive, predizer o valor genético de indivíduos não observados. Demonstra, então, a importância desse tipo de abordagem no caso de experimentos com poucas ou nenhuma repetição por tratamento, como é o caso dos delineamentos aumentados.

Vale relembrar que as informações específicas de parentesco entre as linhagens não foram levadas em conta no presente estudo, mas somente o fato de se relacionarem a uma população de referência comum. Contudo, estas informações podem ser incorporadas à análise por meio dos coeficientes de parentesco, obtidos a partir da genealogia do material (Jiménez & Villa, 1995). Também podem ser utilizadas informações de similaridade genética obtidas por marcadores moleculares (André, 1999). Assim, à semelhança do chamado *modelo animal* (Henderson, 1984), constrói-se uma matriz de parentesco (**A**) que pondera os componentes de variâncias-covariâncias dos genótipos (progênies), em **G**. Neste sentido, também foi bastante simplificada a estrutura aqui assumida para a matriz **G**, isto é, $\mathbf{G}=\mathbf{I}\sigma_g^2$; a qual ignora, além dos parentescos (elementos fora da diagonal principal), as origens diferenciadas das progênies (heterogeneidade das variâncias na diagonal principal), fato este abordado na próxima seção. Entretanto, se o objetivo central da análise estatística é a seleção de genótipos, sempre que for possível deve-se considerar estruturas mais complexas tanto para **G** como para **R**.

A análise teórica apresentada nos parágrafos anteriores, por certo, não traz respostas definitivas para a questão de decidir se os efeitos de blocos e/ou de genótipos são fixos ou aleatórios. Mas, a despeito de alguma indefinição, pode-se afirmar que essa decisão não pode passar simplesmente pela sutileza subjetiva de o pesquisador desejar ou não fazer generalizações a partir do seu experimento. É necessário planejá-lo conforme esses interesses, implantando-o e

conduzindo-o à luz das suposições previamente estabelecidas. Ademais, assumir que os erros experimentais são independentes e homocedásticos ($\mathbf{R}=\mathbf{I}\sigma_e^2$) não pode ser apenas um artifício para garantir conveniência à análise. Mas, passa pela obediência ao princípio experimental de distribuição aleatória dos tratamentos e de seus conjuntos, bem como pela avaliação posterior da estrutura de erros. Assim, não se pode, indiscriminadamente, adotar uma análise interblocos, como não se pode, por simples conveniência, assumir que os genótipos são aleatórios para usufruir das propriedades vantajosas dos preditores *BLUP*'s. Em outros termos, entende-se que não se deve, simplesmente, buscar a análise capaz de captar diferenças cada vez menores, mas aquela que o experimento realmente permite realizar, conforme o seu planejamento, instalação e condução. "O melhor modelo não é necessariamente o que produz os menores erros padrão" (Littell *et al*., 1996). (Modelos fixos, por exemplo, subestimam σ_e^2 sob autocorrelação positiva nos resíduos – Gujarati, 1992). Enfim, a análise não pode resolver todo o problema, também porque não corrige deficiências da amostragem. A estratégia mais eficiente e mais coerente com a realidade é, portanto, planejar e executar os experimentos em sintonia com as suposições que garantam a recuperação do máximo de informações válidas.

3.4. Os modelos de análise sob progênies de diferentes origens

Nos ensaios em que os novos tratamentos têm origens ou procedências diferentes (ex: com progênies de vários cruzamentos) é de interesse que a análise estatística também leve em conta esta informação. No programa de melhoramento de soja desenvolvido no Departamento de Genética da ESALQ/USP este fato realmente ocorre, haja vista seu enfoque de seleção recorrente. Em casos como este o conjunto das progênies, tratado até o momento como uma população única, é subdividido em *c* subconjuntos ou populações, conforme o número de cruzamentos (procedências) dos quais são originárias. Um modelo similar ao proposto por Scott & Milliken (1993) permite incorporar esse tipo de situação:

$$Y_{ijk} = \mu + \beta_j + C_k + g_{i(k)} + \varepsilon_{ij} \qquad \text{(IV)}$$

em que, além de termos anteriormente definidos, tem-se:

C_k : é o efeito fixo do cruzamento *k*, incluindo-se testemunha *(k=1,2,...,c,c+1,c+2,...,c+t)*; e

$g_{i(k)}$: é o efeito do genótipo (progênie ou testemunha) *i* oriundo do cruzamento *k (i=1,2,...,p_k*; p_k é o número de genótipos no cruzamento *k)*, assumido fixo se *i* for uma testemunha, ou aleatório com distribuição $N(0, \sigma^2_{g_k})$ independente, se *i* for uma progênie relacionada ao cruzamento *k* ($\sum_{k=1}^{c} p_k = p$, o número total de progênies).

O fator *C* (aqui denotado '**CRUZ**') compreende, então, *(c+t)* níveis correspondentes às populações fixas sob teste: os *c* cruzamentos que realmente originaram progênies, com variação dentro deles, e as *t* testemunhas, sem essa variação dentro. Assim, a fonte de variação associada aos efeitos $g_{i(k)}$ quantifica exatamente a variação média de progênies dentro dos cruzamentos.

Para a construção de um programa *SAS* que realize as análises fundamentadas neste modelo, é necessário substituir a variável '**TIPO**' no arquivo de dados (**ARQ_ORIG**), pela nova variável de classificação '**CRUZ**'. Se os cruzamentos forem identificados por números, sugere-se, para a compatibilidade com as recomendações de análise que se seguem, que o usuário o faça conforme o índice *k* apresentado anteriormente, ou seja, com as testemunhas correspondendo aos últimos níveis de '**CRUZ**'. Assim, um conjunto de instruções que permite obter a análise intrablocos é:

```
proc glm data=ARQ_ORIG;
  class BLOCO CRUZ GENOT;
  model PG=BLOCO CRUZ GENOT(CRUZ);
  lsmeans CRUZ/stderr cl;
  lsmeans GENOT(CRUZ)/slice=CRUZ stderr cl;
run;
```

A opção '**slice**' permite obter o desdobramento referente à variação de genótipos (progênies) dentro de cada cruzamento. Além disso, apesar de as *t* testemunhas também estarem incluídas na fonte de variação **CRUZ**, não há *GL* dentro de cada testemunha, assim como dentro de um cruzamento com uma só progênie. Portanto, toda a variação em **GENOT(CRUZ)** está relacionada às progênies, não recebendo qualquer influência das testemunhas. O mesmo verifica-se para o seu desdobramento dentro de cada cruzamento (**slice=CRUZ**). O quadro de *ANOVA* correspondente pode, então, ser apresentado como:

F.V.	GL	QM
BLOCO	$b-1$	QM_B
CRUZ	$c+t-1$	QM_C
GENOT(CRUZ)	$\sum_{k=1}^{c}(p_k-1)=p-c$	$QM_{G(C)}$
GENOT(CRUZ*1*)	p_1-1	$QM_{G(C1)}$
GENOT(CRUZ*2*)	p_2-1	$QM_{G(C2)}$
...	...	...
GENOT(CRUZ*c*)	p_c-1	$QM_{G(Cc)}$
Erro	$n-(b+v-2)$	QM_{Erro}
Total	$n-1$	---

Os testes relacionados às variações nos efeitos fixos de 'Testemunhas' ($t-1$ *GL*), 'Cruzamentos com progênies' ($c-1$ *GL*) e 'Testemunhas *vs* Cruzamentos com progênies' (*1 GL*), no

SAS, só podem ser obtidos pelo comando '**contrast**'. Para isso é necessário construir as matrizes específicas, com os coeficientes dos referidos contrastes, conforme o número de testemunhas (*t*) e de cruzamentos com progênies (*c*). Por exemplo, num ensaio com *c=4* e *t=3*, isso pode ser obtido acrescentando-se a instrução a seguir, logo após a linha '**model** ...' do programa anterior:

```
contrast 'TESTEMUNHAS '  CRUZ 0 0 0 0 1 -1 0,
                         CRUZ 0 0 0 0 0 1 -1;
contrast 'CRUZ C/ PROG'  CRUZ 1 -1 0 0 0 0 0,
                         CRUZ 0 1 -1 0 0 0 0,
                         CRUZ 0 0 1 -1 0 0 0;
contrast 'TEST vs CRUZ'  CRUZ 3 3 3 3 -4 -4 -4/divisor=4;
```

Também aqui, no caso de experimentos muito grandes, pode ser conveniente executar a análise através do *PROC MIXED*, ao invés do *PROC GLM*, visando alguma economia no tempo de processamento. A seqüência de comandos também continua muito semelhante, bastando substituir '**proc glm**' por '**proc mixed**' e excluir o termo '**stderr**' das linhas do comando '**lsmeans**'.

Numa análise com recuperação da informação interblocos, ainda com progênies de efeitos fixos, também é conveniente substituir o *PROC GLM* pelo *PROC MIXED* para garantir saídas corretas de erros de estimativas (o procedimento *GLM* continua útil para obter quadros de *ANOVA* e as expressões de *E(QM)*). No caso desta análise, um conjunto de instruções que permite a sua realização é:

```
proc mixed data=ARQ_ORIG;
  class BLOCO CRUZ GENOT;
  model PG=CRUZ GENOT(CRUZ);
  random BLOCO;
  lsmeans CRUZ/cl;
  lsmeans GENOT(CRUZ)/slice=CRUZ cl;
run;
```

Para os modelos com progênies aleatórias (modelos mistos *B* e *C*) são necessárias algumas adaptações adicionais. Primeiramente, deve-se gerar, no arquivo de dados (**ARQ_ORIG**), *c* variáveis auxiliares, aqui denominadas **NEW*1***, **NEW*2***, ..., **NEW*c***, construídas com princípio idêntico à variável **NEW** já descrita na seção *2.3*. Assim, **NEW*1*** é igual a *um* se as progênies forem do **CRUZ*1*** e a *zero* em caso contrário; **NEW*2*** é igual a *um* se as progênies forem do **CRUZ*2*** e a *zero* em caso contrário; e assim por diante. Para isso, deve-se introduzir as seguintes instruções no programa de leitura de dados (item *2.3.1*), imediatamente antes da linha "**run;**" (admita, por exemplo, *c=4* cruzamentos):

```
if CRUZ=1 then NEW1=1; else NEW1=0;
if CRUZ=2 then NEW2=1; else NEW2=0;
if CRUZ=3 then NEW3=1; else NEW3=0;
if CRUZ=4 then NEW4=1; else NEW4=0;
```

Após a execução do programa de leitura de dados, as análises podem ser executadas com as respectivas seqüências de comandos (exemplos com *4* cruzamentos):

Modelo misto B (blocos fixos e progênies aleatórias):

```
proc mixed data=ARQ_ORIG method=reml;
 class BLOCO CRUZ GENOT;
 model PG=BLOCO CRUZ/ddfm=satterth;
 random GENOT*NEW1 GENOT*NEW2 GENOT*NEW3 GENOT*NEW4/solution cl;
 lsmeans CRUZ/cl;
run;
```

Modelo misto C (blocos e progênies aleatórios):

```
proc mixed data=ARQ_ORIG;
 class BLOCO CRUZ GENOT;
 model PG=CRUZ/ddfm=satterth;
 random BLOCO GENOT*NEW1 GENOT*NEW2 GENOT*NEW3 GENOT*NEW4/
   solution cl;
 lsmeans CRUZ/cl;
run;
```

Nestes dois programas, o comando '**lsmeans**' (para efeitos fixos) lista as médias ajustadas dos cruzamentos e das testemunhas ($\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_c, \hat{\mu}_{c+1}, \hat{\mu}_{c+2}, ..., \hat{\mu}_{c+t}$), permitindo identificar as populações mais promissoras em valor genético médio. Os componentes de variância estimados por *REML* (outros métodos são disponíveis) e para cada um dos cruzamentos ($\hat{\sigma}^2_{g_1}, \hat{\sigma}^2_{g_2}, ..., \hat{\sigma}^2_{g_c}$) permitem avaliá-los, individualmente, em termos de variabilidade genética. Assim, combinando-se as informações (médias e variâncias) é possível avaliar o potencial de cada cruzamento para produzir genótipos (cultivares) promissores. Deve-se esclarecer que, nestes modelos, os preditores $\widetilde{g}_{i(k)}$ (*EBLUP*'s listados pela opção '**solution**') são comparáveis apenas dentro de cada cruzamento (população de referência, com μ_k e $\sigma^2_{g_k}$ comuns). Assim, a seleção de genótipos deve seguir uma das alternativas: *i)* seleção entre e dentro de cruzamentos, escolhendo-se primeiramente entre os cruzamentos (com base em suas médias) e, em seguida, entre as progênies dentro dos melhores cruzamentos (com base nos preditores); ou, *ii)* seleção global envolvendo as progênies de todos os cruzamentos, com base nas médias preditas das progênies, as quais são obtidas por: $EBLUP(\mu_k + g_{i(k)}) = \hat{\mu}_k + \widetilde{g}_{i(k)}$ (sendo: $k=1,2,...,c$ e $i=1,2,..., p_k$).

Apesar da aparente simplicidade na implementação da análise (com progênies de diferentes origens), no caso de progênies aleatórias (modelos mistos *B* e *C*), alguns problemas podem inviabilizar a sua aplicação. O primeiro deles é que a análise só terá utilidade prática se os cruzamentos, individualmente, apresentarem um considerável número de progênies dentro deles.

Caso contrário, as estimativas de $\sigma^2_{g_k}$ terão precisão muito baixa, assim como os preditores dos valores genotípicos individuais (*EBLUP*'s), o que reduz sensivelmente a eficiência da análise. Ademais, o tempo de processamento da análise eleva-se consideravelmente para um número de cruzamentos ainda relativamente baixo (ex: *c=10*), prejudicando a sua aplicação rotineira, pelo menos, nos dias atuais. Assim, com freqüência, surgem também problemas relacionados à insuficiência de memória computacional. Isto porque, nessa abordagem (utilizando as variáveis **GENOT*NEW***k*, com *k=1,2,...c*), o número de colunas da matriz **Z**, de incidência dos efeitos aleatórios no correspondente modelo misto, torna-se igual a *vc*, ao invés de simplesmente *v*, do enfoque principal dado neste artigo (com progênies de uma só procedência).

O consumo de memória computacional, em *bytes*, pelo *PROC MIXED* do *SAS* é função direta da dimensão das matrizes $\mathbf{X}_{(nxp)}$ e $\mathbf{Z}_{(nxq)}$ no modelo misto: $40(p^2+q^2)+32(p+q)^2$ (SAS Institute, 1997). Entre as alternativas que permitem alcançar alguma economia de memória e redução no tempo gasto em CPU (Unidade Central de Processamento), a primeira seria listar no comando '**model**' apenas os produtos **GENOT*NEW***k* (*k=1,2,...,c*) cujos cruzamentos apresentem duas ou mais progênies. Isto porque os demais, embora aumentem a dimensão de **Z**, terão $\hat{\sigma}^2_{g_k}$ e *EBLUP*'s obviamente nulos. O SAS Institute (1997) faz outras recomendações, sob insuficiência de memória e/ou tempo excessivo de processamento durante a aplicação do *PROC MIXED*. Entre estas, aplicam-se neste caso: o uso de '**mivque0**' (não iterativo), em substituição a '**reml**' (iterativo), como método de estimação dos componentes de variância ('**method=**'); e o uso da opção '**ddfm=bw**' (de uso padrão em análise de medidas repetidas), ao invés de '**ddfm=satterth**', para a aproximação dos *GL* de denominadores para as estatísticas *F* e *t*.

Para a seleção entre cruzamentos e a avaliação da variabilidade genética dentro deles, outra alternativa é utilizar os resultados da análise intrablocos. Para isso é necessário obter as estimativas *ANOVA* de $\sigma^2_{g_k}$, pelo método dos momentos, conforme aplicação de Hamawaki (1998). O estimador correspondente é: $\hat{\sigma}^2_{g_k} = (QM_{G(Ck)} - QM_{Erro}) / \mathrm{K}$; sendo K o coeficiente de $\sigma^2_{g_k}$ na expressão: $E(QM_{G(Ck)}) = \sigma^2_e + \mathrm{K}\sigma^2_{g_k}$. O valor de K, uma função de atributos de tamanho no experimento (p_k, *b*, etc.) pode ser obtido através da função '**contrast**' associada ao comando '**random**', via *PROC GLM* do *SAS*. É evidente que, adotando-se esta abordagem, não se obtém diretamente a predição dos valores genotípicos individuais (*EBLUP*'s) das progênies. Embora isto possa ser obtido pelo *PROC MIXED*, fornecendo-se as estimativas *ANOVA* dos componentes de

variância através do comando '**parms**', associado às opções '**noiter**' e '**noprofile**' (Federer & Wolfinger, 1998; Littell *et al.*, 1996).

4. CONCLUSÕES

A avaliação das diferentes alternativas de análise estatística para blocos aumentados permitiu concluir que os modelos que recuperam informações inter-efeitos (interblocos ou/e intergenotípica), em geral, garantem médias de tratamentos mais precisas, particularmente, se a informação aproveitada for de natureza genotípica (intertratamentos). Ademais, embora haja grande concordância entre os métodos avaliados (*Modelos 0*, *1*, *2*, *3*, *4* e *5*) quanto ao ordenamento dos genótipos, se a herdabilidade for baixa ($h^2_{Y_{ij}} < 0,20$), os valores das respostas genotípicas médias poderão diferir substancialmente entre os grupos de modelos (*0*, *1*, *2*, *5*) e (*3*, *4*). Isto resulta do fato de que os modelos de análise que recuperam informação interprogênies (*3* e *4*), quase sempre, produzem médias mais uniformemente distribuídas em torno da média geral de progênies do que os demais. Em função disso, normalmente espera-se que uma menor proporção de genótipos supere a melhor testemunha (em produtividade média de grãos) nas análises obtidas a partir dos dois modelos. Assim, sobretudo em condições de variabilidade genética realmente baixa (ex: espécies já bastante melhoradas) e de precisão experimental deficiente (ex: delineamentos aumentados), o uso de modelos que desconsideram tal informação pode implicar em desperdício de tempo e de recursos com a avaliação de material genético pouco promissor, mantido nos próximos ciclos seletivos. No contexto atual do Programa de Melhoramento da Soja desenvolvido na ESALQ/USP, o modelo de análise que melhor se adaptou às características dos ensaios em blocos aumentados foi o de blocos fixos e tratamentos adicionais aleatórios (*3*).

Conclui-se também que genótipos bem classificados por alguns procedimentos estatísticos são nitidamente descartados por outros, indicando que uma escolha equivocada do método de análise pode comprometer a eficiência do programa de melhoramento. Neste sentido, a seleção com base em testemunhas intercalares (*Modelo 5*) apresentou a menor coincidência de genótipos selecionados com os outros modelos, exceto com a análise intrablocos. Enquanto as maiores concordâncias ocorreram entre os modelos: *0* e *2*; *1* e *2*; e, *3* e *4*. Enfim, pôde-se constatar a importância de uma especificação apropriada do modelo de análise para o resultado final da seleção, haja vista as conseqüências da simples mudança na suposição dos efeitos de novos genótipos, se fixos ou aleatórios, sobre suas classificações e respostas fenotípicas esperadas, sobretudo em relação aos padrões comerciais (cultivares testemunhas).